The Geographic Distribution of Human Capital in China

Jiang, Ying

Master thesis for the Master of Philosophy in

Environmental and Development Economics

UNIVERSITETET I OSLO

May 2011

Preface

Finally I finished this thesis and it is the time to say goodbye to the wonderful two-years in University of Oslo. It is also the time to say thanks to some special people. Because of them, I am able to finish the thesis, graduately understand the beauty of economics and fall in love with it.

My deepest gratitude goes first and foremost to my supervisor Finn Ragnar Førsund. He shows great interests in the topic which he is not very familiar with. And he always encouraged me to develop original ideas, gave me exciting inspirations and provided insightful suggestions. I am also grateful for his deep thoughts about the relationship between theoretical studies and empirical studies. The freedom he gave me to move on is absolutely invaluable.

I would also like to thank Arne Strøm, who is so considerate, patient and earnest, and makes mathematics a real tool for economics. You are the loveliest professor I have ever met. And also thanks to Geir B. Asheim, Kjell Arne Brekke and Tapas Kundu for the wonderful lectures and seminars in Microeconomics. Thanks to Tore Schweder for his passionate lecturing and rich experiences in econometric study. Thanks to all the other lecturers I have met in university of Oslo. Without their interesting illustrations and careful explanations, I could not have so much fun in all those figures and models.

I am also thankful to those friends who shared experiences and knowledge during my study, who make winter in Oslo so warm.

Finally, I dedicate this thesis to my parents, who gave me enormous encouragement to study abroad and try to be excellent. My profound gratitude goes to my loving, supportive, encouraging, and patient husband Ran Gao, who supported me academically and emotionally through the rough road to finish the thesis. Thank you.

May 2011

Jiang, Ying

Summary

In this paper, we basically made efforts to answer 3 questions: （1）What is the geographic distribution of human capital in China? (2) How the regional differences of human capital change over time? (3) What are the main determinants of geographic distribution of human capital?

To answer the first question, we choose the education attainment approach and build an indicator system which includes both a human capital stock indicator and human capital structure indicators. Average years of schooling is the human capital stock indicator, and percentage of people with higher education, high school education, middle school education, primary school education (according to the final education level) and illiteracy rate are the 5 human capital structure indicators. We collect data of 31 regions from 1997 to 2008, and run the cluster analysis. We find that Beijing, Tianjin, Shanghai are most developed in human capital, and Tibet is ranked the last. Furthermore, northern regions and some central regions are better developed in human capital; western regions are generally poor in human capital. However, not all economically developed regions have an abundant educated population, like Zhenjiang and Fujian; and not all poor regions lack human capital, like Shanxi.

To answer the second question, we compute the correlated variation of each indicator in the observation period. Through comparison among indicators, we find that percentage of people with higher education varies most in 31 regions, then the illiteracy rate, then the percentage of people with high school education. The correlated variations of the rest indicators are relatively small. Meanwhile, the regional difference of illiteracy rate tends to increase during the 12 years, and so does the regional difference of percentage of people with primary school education. And these two indicators represent the situation of basic education. Regional differences of the rest decline. Since it is unreasonable to have a high regional difference in illiteracy rate, we compute the results by excluding Tibet again, and find that this change does not affect other indicators but do decrease the correlated variation of illiteracy rate. But the increasing trend does not change.

To answer the third question, we do two things. First, we run two panel data regressions. The dependent variable of first one is average years of schooling, and of the second one is

percentage of the population with higher education. The model we use is the fixed effect model including both entity and time fixed effects. Nine regressors are added in and two of them are control variables. We find that income, the number of health personnel per 10,000 inhabitants and the number of street lights per cities are positively related with the two dependent variables. The last two are variables reflecting nonpecuniary benefits offered by the region. Cost has a negative impact on average years of schooling, but does not affect percentage of people with higher education. Government's education expenditure per person does increase average years of schooling in a statistical perspective, but government's expenditure on higher education per student does not affect the percentage of people with higher education. The differences in the results of the two regressions may suggest different behavior patterns among people with different education backgrounds. The unemployment rate is not statistically significant in either of the two regressions. Two possible reasons could explain this: first the data we used is not the unemployment rate of educated people, but the registered unemployment rate of urban residents, because the first one is not available. Secondly, lack of full information may also lead to this result. The other thing we do to answer question (3) is constructing an internal migration model to explain the formation of human capital in labor force. In this model, people possess different levels of latent ability, and have to make two decisions: migration for education and migration for job. Furthermore, people assign different weights to nonpecuniary benefits. As a result, education cost, living cost is negatively related with human capital stock in that region, on the contrary, starting wage offered by employers and nonpecuniary benefits are positively related with human capital stock in that region. A brief discussion of preferences is also made.

Contents

1 Introduction

Regional gap is an inevitable consequence for developing countries when they experience rapid economic transformation and growth. China went on the track of sound economic progress after Xiaoping Deng stressed his famous point “let some people become rich first”. However, unbalanced development among regions comes as a consequence of the rapid growth in the whole nation, and it exists in many aspects.

It is estimated that there is an "U-shaped” pattern in regional inequality during the reform period (1978-1994), measured by the provincial per capita GDP. The diminishing trend is due to a significant decline in income inequality among provinces in the coastal regions, and on the other hand, the rising GDP in the southern coastal belt also leads to widening regional disparity between costal and internal regions (Ying, 1999). At the same time, the income gap between rural areas and cities also appeared, and caused a large-scale rural-to-urban migration (Zhang and Meng, 2010). This migration trend partially contributes to the segmentation in urban labor market. Most migrants are not granted urban citizenship and treated as “outsiders”; while a few of them who can be considered as “elites” are able to obtain permanent resident permit; It is found hat in terms of human capital attributes, mobility resources, and labor market entry and shifts, permanent migrants are the most privileged and successful elite, followed by non-migrant natives, and finally by temporary migrants at the bottom of the hierarchy (Fan, 2002). However, rural migrants can not obtain wage comparable to their urban counterparts in their life time, and most importantly, even well-educated rural migrants do not seem to have a significant advantage comparing to the poorly-educated ones (Zhang and Meng et.al, 2009).

In such a scenario, with widening regional income gap, large scale of migration trend and segmentation of labor market, different regions exhibit different ability in human capital accumulation. However, the unbalanced development in human capital seems not a very attractive academic topic to researchers. On the other hand, the fact of unbalanced development in human capital among regions has been well noticed. In western regions of China, drop-out rate from schooling is higher than average, especially in rural areas (Li, 2010).

For some poor regions, most of which are located in the western part, it is even difficult to find qualified teachers or health personnel, because of a serious out-flow of educated people (Yang, 2010). And college students from different regions gather in big cities like Beijing, Shanghai and Guangzhou. All of these facts contribute to the unequal geographic distribution of human capital. Indeed, to some extent, the unequal distribution of human capital has led to the failure of Xiaoping Deng's another principle, which is "Material civilization and spiritual civilization should run neck and neck."

Therefore, there are actually sound academic and practical ground for a careful research on the geographic distribution of human capital and its determinants. On one hand, the effect of human capital on economic growth has been proved to be positive both by new growth theory and empirical studies in China (Fleisher et al., 2010). On the other hand, human capital accumulation is a main theme of the human development for individuals, and also a major element of "spiritual civilization". Thus, if we can picture a clear map of the geographic distribution of human capital, and find out the underlying determinants, it will help to balance the development among regions.

There are some attempts to describe the unequal distribution of human capital in China. And a common approach is to calculate the average years of schooling. Because of the data limitation, the average years of schooling in labor force is not available, only the average years of schooling among people aged over six can be calculated. Furthermore, some studies divided regions into several groups according to geographic locations, such as northern part, eastern part, central part and western part, and studied how each group differs in the average years of schooling, government's education expenditure and etc (Liu et al., 2008).Some studies used cluster analysis and categorized regions into several clusters. The first approach focus on the geographic location, but regions in the same group may differ a lot in human capital stock. The second approach is more helpful, but in previous studies, the variables used in cluster analysis actually include both economic indicators and human capital indicators. For example, some scholars selected per capital GDP and the population ratio of people working in agriculture as variables and put them in the cluster analysis of human capital (Xu and Wang, 2006). This mix-up of indicators could not reflect the pure difference of human capital exactly. Thus, building a new indicator system is necessary, which only includes

human capital indicators and reflects the situations of both human capital stock and human capital structure.

On the other hand, it is of our interests to see how regional difference in human capital has changed with time. There are not many relevant researches. Chen, Zhao has studied the evolution of regional difference in human capital during 1987-2001 through comparing the correlated variance of some selected indicators (Chen et al., 2004). However, the observation period is relatively far from present. Therefore, to examine the recent trend of regional difference in human capital is another task of this paper. We use the data from 1997 to 2008, compute the correlated variation of all the 6 indicators, and make comparison among indicators in the same year and among different years for the same indicator.

It is also of great importance to understand how related characteristics in a certain region affect the human capital stock there, since it will help local government to make well-targeted policies. However, related empirical study is rare. Furthermore, even for the existing researches, there are still some improvements we can make. For example, Yang, Yang collected the data of 2008 and checked several potential factors which may affect human capital difference among regions (Yang, 2009). But regressions based on one year's data could not eliminate omitted variable bias, which is resulted from failing to include some regional and time effects. In this case, panel data regression is an useful tool (Stock and Watson, 2007), and is also the approach we will adopt in this paper. The panel data covers 31 regions and includes 11 years. Dependent variables are average years of schooling and the percentage of people with higher education. Independent variables are those which have economic implications, such as income, cost, unemployment rate and etc. Panel data regression enables to reduce certain types of omitted variables, but not all of them. As a result, two control variables are added: the first one reflects the student ratio in the whole population, which is positively related with both income and the dependent variable; the second one reflects the population structure, since it is the average year of schooling among population aged over six we have used, population structure has a significant impact on the dependent variable, and also correlated with student ratio. Population structure is also correlated the average annual income as well as expenditures since different age groups have different earning abilities and consumptional habits. Despite of the efforts above, validity problems

may still exist, especially the simultaneous causality problem, thus we also make a discussion of them and further improvement is appreciated.

As we mentioned before, due to the data limitation, we can only focus on the human capital embedded in the population aged over six, and could not examine the human capital embedded in labor force. Consequently, a theoretical model may help to further our understanding of the human capital difference in labor force among regions. In fact, previous papers in human capital investment and international brain drain provided some inspirations, but are not able to answer the question directly. Because two process contribute to the formation of human capital in labor force in a region: migration for education and migration for jobs. The second one is often discussed in papers about international migration, but the first one is not, due to the differences in settings between international migration and internal migration. Thus, we will incorporate the two processes and construct an internal migration model to explain the formation of human capital in labor force in a certain region.

Based on the discussion above, the rest of the thesis is organized as follows: in chapter 2, we discuss the approach to measure human capital stock as well as human capital structure, and then select proper indicators. Thereafter, necessary data is collected; a cluster analysis based on the data is done and the 31 regions will be categorized into several types according to their situations of human capital stock and structure. In chapter 3, we examine how regional differences in human capital have changed during 1997-2008. In chapter 4, we run the panel data regression against to two dependent variables: one is average years of schooling; the other is percentage of people with higher education. In chapter 5, an internal migration model is given, and the mechanisms which affect human capital formation in labor force are revealed. In the last chapter, based on the conclusion of this paper some suggestions are made.

2 Geographic distribution of human capital in China

2.1 Measurement

Adam Smith defined four fixed capitals in production: machines, buildings, land, and the last one is “the acquired and useful abilities of all the inhabitants or members of the society” (Smith, 1776). This can be considered as early understanding of human capital. “Human capital” was first used as a term in an article by Arthur Cecil Pigou in 1928, and was developed to a modern theory under the efforts of Jacob Mincer, Gary Becker and T. W. Schultz at the University of Chicago in 1960s. According to Becker, “Activities that influence future monetary and psychic income by increasing the resources in people [...] are called investments in human capital.” (Becker,1964). Although this definition has been extended or modified by researchers later, the three major issues related to human capital has not changed:

1) Human capital is embodied in individuals:
2) Human capital was formed by past investment;
3) Human capital influences future income.

There are also the three principles to follow when discussing the measurements of human capital. Thus, we can categorize approaches of measuring human capital into 3 types: the cost-based approach, the income-based approach, and the educational attainment approach. To select a proper approach to measure human capital, we should have a careful comparison of the three approaches first.

The cost-based approach: Some scholars also call it “the retrospective approach”. This approach estimates the past cost of “producing” an individual, including both tangible and intangible costs, such as nutrition, expenditure on health, education and training. A problem associated with this approach is how to identify each kind of cost clearly. The other drawback is it ignored the social cost which also contributes to the accumulation of individual’s knowledge and skills, such as government education subsidy, etc.

The income-based approach: Since human capital is a factor of producing economic value, another way of measuring it is to estimate the present value of the future benefits. This

approach has a higher requirement of data, for example, earnings and employment rate by education, which is not usually available in developing countries.

The educational attainment approach: Jacob Mincer, Gary Becker and T. W. Schultz modeled human capital investment as individuals' rational choice in their papers in the middle of 20th century. This also led to a new approach of measuring the quantity of human capital; that is using several education-related indexes, including average years of schooling, the ratio of people with certain level of education, the student-teacher ratio, etc. Those approaches have some drawbacks. For example, average years of schooling disregards the quality differences among each education level, and also failed to take on-job-training into account. However, this approach has a relatively lower requirement in data, and that is why it was used by many researchers in China.

Keeping all those advantages and disadvantages in mind, and taking data availability into account, we finally choose to follow the third approach: the educational attainment approach.

Now we have the proper approach, and there are still two questions to be answered before we start the data collection. The first question is "human capital of which group". As we mentioned before, most papers and publications in this field are about the relationship between human capital and economic growth. And in this scenario, "human capital" is the human capital in workforce, which has direct connection with production. However, we do not have relevant data to compute the human capital in workforce in China. Various yearbooks only provide the information of educational attainment in people aged over six, which includes both labor force and children. Thus, it is impossible to distinguish labor force from children based on this information. Besides, this method has its own disadvantages. First, it is difficult to draw a simple line between workforce and non-workforce according to age. Some scholars use the data of people who are above 15, and some think it should be above 25. In fact, this disadvantage becomes a more serious problem when we study the situation in developing countries, because some children drop out from school very early and work illegally. Furthermore, our study here is about human capital itself, not the relationship with other variables. Thus, it is reasonable to take into account the human capital of all people who are eligible for education, not solely the human capital of workforce. Based on discussion above, we finally choose people aged above six as the target group in empirical study.

Nevertheless, we will use a separate chapter to discuss the determinants of geographic distribution of human capital in labor force, by constructing a theoretical model.

The second question is what kind of indicators we should use to describe the human capital. On one hand, human capital is a concept of "stock", so it is very nature to use some stock indicators. And on the other hand, indicators of stock usually disregard some fundamental characteristics of the variable, like structure. So besides the indicators of stock, we also need some other indicators which will help us to have a deeper insight of human capital. Since we use average years of schooling as the stock indicator, the selection of other indicators becomes easier: following the educational attainment approach, we can obtain several other indicators, which is able to reflect the structure of human capital in a certain region. These indicators are: percentages of population holding their highest education level as college or above, high school, middle school and primary school respectively, and the illiteracy rate.

Table 2.1 Summary of the indicators

INDICATORS	FORMULA	EXPLAINATION
Average years of schooling	$AYS_t = \frac{\sum N_{it} \cdot Y_i}{N_t^{6+}}$	N_{it} : the number of people with i level of education in year t (i=c, college; i=h high school; i=m middle school; i=p primary school; i=i illiteracy)
Percentage of college	$POC_t = \frac{N_{ct}}{N_t^{6+}}$	
Percentage of high school	$POH_t = \frac{N_{ht}}{N_t^{6+}}$	Y_i : years of schooling related to i level of education
Percentage of middle school	$POM_t = \frac{N_{mt}}{N_t^{6+}}$	N_t^{6+} : the number of people who aged over 6;
Percentage of primary school	$POP_t = \frac{N_{pt}}{N_t^{6+}}$	N_t^{15+} :the number of people who aged over 15
Illiteracy rate	$IC_t = \frac{N_{it}}{N_t^{15+}}$	

Till now, we have constructed an indicator system to evaluate human capital in a certain region. And to make things simple, in the following discussions we will call the five structure indicators as: percentage of college, percentage of high school, percentage of middle school,

percentage of primary school and illiteracy rate. A more specific description is shown in table 2.1.

It is worth mention that the 5 structural indicators are not summed to 1, because illiteracy rate here is the rate among people who aged over 15, not six. And the reason why we use illiteracy rate among people aged over 15 is simply because this data is available in relevant yearbooks.

2.2 Data description

With the indicators discussed above, we can now have a look at the geographic distribution of human capital in China. China's Statistic Yearbook provides each region's illiteracy rate and number of people according to different education levels in the previous year, which is the basic data for computing the values of indictors. We will compute the values of the indicators for all the 31 regions in China, including 22 provinces, 5 autonomous regions, 4 municipalities. The two Special Administrative Regions: Hong Kong and Macau are not included, and Taiwan is also not included. According to the formulas in table 2.1, we need the information about the duration of each level of education. However, not every region has the same education system. For example, primary schools in Shanghai and Hubei Province require only 5-years of study, but in most other regions, the requirement is 6 years. Furthermore, even in the same region, years of schooling for the same education level may differ. So this makes the determination of Y_i becomes difficult. The Department of Population and Employment Statistics in National Bureau of Statistics of China suggested that: relevant years of schooling for college, high school, middle school and primary school can be calculated as 16, 12, 9 and 6 respectively. And people who attended the class for eliminating illiteracy can be considered as having accepted 1 year of education. We will follow this suggestion when computing average years of schooling.

11 years data are all available in each year's "China's Statistic Yearbook": from 1997 to 2008, excerpt data in 2001. Yearbooks before 1997 do not provide any information about number of people according to different education levels for each region. And China's Statistic Yearbook in 2002 did not report the relevant data in 2001; instead, it adjusted the data obtained through 2000's Census, which has already been revealed in 2001's yearbook. Thus we choose the adjusted data to describe situation in 2000, and no data is available for year 2001.

Table 2.2 Summary of the data: AYS, POC, POH, POM, POP, IC

	1997						1999						2002					
	ASY	POC	POH	POM	POP	IC	ASY	POC	POH	POM	POP	IC	ASY	POC	POH	POM	POP	IC
Beijing	8.04	13.44	25.38	34.74	19.67	7.64	9.98	19.13	23.56	33.49	17.99	6.45	10.26	20.49	23.99	35.67	14.86	5.35
Tianjin	6.97	7.27	19.07	34.72	29.98	9.84	8.71	7.90	21.28	36.32	27.08	8.03	9.15	10.57	22.57	37.36	23.16	6.74
Hebei	5.78	2.10	10.81	35.33	39.35	14.30	7.46	2.82	10.68	38.16	38.26	11.42	8.03	4.69	11.61	42.50	34.41	7.81
Shanxi	6.13	3.36	11.69	38.77	37.54	9.87	7.82	3.82	11.71	41.85	34.02	9.14	8.25	4.63	12.67	45.41	31.63	6.40
Inner Mongolia	5.94	3.60	12.04	32.49	37.29	16.78	7.35	3.80	13.88	33.90	33.84	16.44	7.88	5.64	14.91	37.82	29.70	13.46
Liaoning	6.42	6.02	12.63	40.05	33.63	8.21	8.18	5.74	12.55	42.25	32.62	7.18	8.44	5.52	13.09	46.68	29.75	5.16
Jilin	6.64	4.99	16.35	33.75	37.17	8.13	8.23	4.93	16.76	37.56	34.15	6.81	8.61	6.50	17.16	39.78	32.23	4.36
Heilongjiang	6.33	4.75	12.76	37.70	36.19	9.18	7.82	3.71	12.13	42.02	33.25	9.77	8.30	4.87	14.71	43.24	31.07	6.54
Shanghai	7.30	8.89	23.82	36.98	21.27	10.17	9.27	11.06	26.04	35.71	19.30	8.68	9.59	15.07	25.10	34.70	17.50	8.18
Jiangsu	5.56	2.05	11.65	34.18	35.15	19.28	7.30	4.02	12.62	34.02	34.71	16.79	7.59	3.83	13.14	38.69	31.96	14.31
Zhejiang	5.57	2.50	10.09	31.58	39.35	18.38	7.14	2.48	11.59	34.84	37.03	15.70	7.68	5.77	13.25	34.54	34.23	13.54
Anhui	5.43	1.83	7.54	32.15	41.11	20.17	6.54	1.62	6.79	34.00	40.17	20.28	6.99	2.64	7.39	38.66	36.65	17.88
Fujian	5.82	2.63	9.38	26.99	45.98	17.45	6.77	2.26	10.07	29.69	42.21	18.46	7.46	4.20	13.15	32.04	38.72	13.67
Jiangxi	5.96	1.88	9.60	31.19	46.46	12.47	7.12	2.44	10.16	31.25	44.90	13.15	7.48	2.91	11.48	34.83	41.68	10.76
Shandong	5.28	1.49	8.60	33.22	37.36	22.64	6.82	1.67	9.46	37.85	33.61	20.15	8.08	5.67	14.42	41.76	28.06	11.24
Henan	5.60	1.60	9.36	38.34	37.87	14.88	7.10	1.90	9.04	40.11	35.08	16.31	8.08	4.30	11.97	46.47	29.51	9.14
Hubei	6.07	3.39	11.58	31.08	41.55	15.05	7.29	3.14	11.33	34.61	38.56	14.98	7.34	3.86	12.23	32.25	39.25	15.13
Hunan	5.92	2.02	10.34	32.66	45.33	11.27	7.45	2.74	11.73	34.43	41.69	11.13	7.91	4.35	12.47	38.69	37.26	8.35
Guangdong	6.23	3.67	11.82	32.25	43.26	9.61	7.61	3.68	12.13	35.02	40.26	9.23	8.09	5.15	13.84	37.78	36.83	7.01
Guangxi	5.49	0.93	6.65	30.54	48.64	15.12	6.84	0.86	7.12	33.45	47.37	12.35	7.62	3.48	11.32	37.03	39.57	9.45
Hainan	5.90	2.32	11.96	33.76	39.45	14.11	7.25	3.66	11.68	33.08	38.12	14.58	7.94	3.59	14.62	39.35	34.55	8.87
Chongqing	5.52	1.97	7.80	27.62	47.75	16.82	6.88	2.28	9.04	29.88	45.61	14.75	7.44	3.35	10.38	34.63	42.33	10.31
Sichuan	5.52	1.99	8.54	27.53	45.83	18.00	6.66	1.95	7.77	30.18	44.92	16.77	7.29	3.75	10.44	33.99	39.58	13.55
Guizhou	5.20	2.01	5.93	21.41	48.17	25.88	6.08	2.26	7.10	23.66	45.56	24.46	6.73	3.52	7.54	30.03	42.71	18.74
Yunnan	5.01	1.24	5.91	21.36	49.31	25.22	5.82	1.19	4.95	22.61	50.00	24.34	6.12	1.99	6.37	25.19	46.15	23.10
Tibet	3.26	0.32	1.32	4.88	47.51	54.08	2.95	0.09	0.32	4.26	41.87	66.18	4.32	0.79	2.87	11.72	46.63	43.82
Shanxi	5.88	2.99	11.89	30.80	39.81	17.34	7.14	3.30	12.51	32.63	36.30	18.29	7.43	3.95	12.99	34.59	35.43	15.56
Gansu	5.18	1.65	9.78	26.11	39.00	26.77	6.35	2.62	10.17	27.06	37.94	25.64	6.78	3.05	11.61	28.86	38.36	21.11
Qinghai	4.07	2.01	6.92	18.31	31.55	43.62	5.97	3.99	10.64	21.79	34.87	30.52	6.35	3.15	8.95	27.65	38.02	24.77
Ningxia	5.33	3.27	10.56	29.01	34.11	25.83	6.66	3.24	10.46	31.26	34.58	23.32	7.39	5.66	11.94	33.60	33.82	17.49
Xinjiang	6.38	5.58	12.53	27.75	43.64	11.52	7.94	7.33	14.18	29.86	39.65	9.77	8.37	9.88	14.85	31.73	35.80	8.21

Table 2.2 Summary of the data: AYS, POC, POH, POM, POP, IC (Continued)

	2004						2006						2008					
	ASY	POC	POH	POM	POP	IC	ASY	POC	POH	POM	POP	IC	ASY	POC	POH	POM	POP	IC
Beijing	10.56	23.89	23.44	33.89	14.55	4.48	10.95	29.36	23.15	29.47	13.70	4.47	10.97	28.12	23.42	31.74	13.39	3.11
Tianjin	9.64	14.34	23.62	36.67	20.26	5.39	9.73	15.22	22.78	36.44	21.34	4.10	9.88	15.46	24.76	36.94	18.48	3.52
Hebei	8.38	5.89	13.15	45.53	29.36	6.76	8.13	3.93	11.43	47.36	31.12	6.42	8.36	4.81	11.99	48.61	29.58	4.83
Shanxi	8.38	5.23	11.51	49.00	29.25	5.75	8.70	6.65	14.60	47.28	27.11	4.42	8.81	7.20	14.77	48.36	25.57	4.24
Inner Mongolia	8.17	6.63	15.06	38.96	29.91	10.46	8.19	6.51	14.72	39.76	30.10	9.36	8.37	7.42	13.91	42.14	28.66	8.14
Liaoning	8.84	8.29	13.65	48.12	25.73	4.36	8.92	9.57	14.65	45.18	26.10	4.12	9.08	11.00	14.21	45.41	25.43	3.48
Jilin	8.80	6.85	17.53	42.98	28.86	3.85	8.66	7.02	16.91	41.98	28.81	5.21	8.89	7.57	17.81	44.74	25.27	4.44
Heilongjiang	8.49	4.68	13.91	48.81	28.03	4.81	8.53	6.11	14.82	44.52	29.49	4.97	8.70	5.97	15.71	47.35	26.65	4.16
Shanghai	10.11	18.50	27.52	32.81	14.97	6.54	10.44	21.83	25.88	33.30	14.06	4.92	10.55	22.66	25.07	34.09	14.05	3.97
Jiangsu	7.81	4.92	13.82	39.57	30.01	13.15	8.25	7.24	15.20	38.62	29.91	9.36	8.44	7.04	15.75	41.95	27.52	8.05
Zhejiang	7.95	7.47	14.60	35.56	30.04	13.40	8.06	8.42	12.91	34.83	33.82	10.20	8.24	9.53	12.55	35.80	33.10	9.38
Anhui	7.49	4.43	10.72	38.76	33.39	15.08	7.34	4.72	9.61	38.51	32.71	16.30	7.44	4.01	10.32	40.15	32.44	14.49
Fujian	7.49	4.56	14.57	32.15	35.34	15.25	7.73	5.83	12.39	34.13	37.23	11.31	7.80	5.86	12.97	34.41	36.88	10.38
Jiangxi	7.98	4.67	14.21	37.72	35.59	9.06	7.71	4.74	11.48	34.46	41.23	9.21	8.26	6.39	14.90	36.40	36.15	6.49
Shandong	7.94	5.49	13.67	42.01	27.41	12.56	8.09	5.73	13.32	41.88	30.16	9.13	8.28	5.47	14.36	44.42	28.00	7.96
Henan	8.22	4.42	13.15	46.83	28.65	8.08	8.05	4.14	11.48	47.98	28.28	8.64	8.34	4.72	13.46	49.36	25.39	7.36
Hubei	8.10	5.78	15.81	38.65	29.92	11.47	8.26	7.71	15.27	37.42	30.40	9.83	8.49	8.08	16.05	39.00	29.28	7.69
Hunan	8.16	5.22	13.47	40.90	33.74	7.44	8.17	5.07	13.93	39.86	34.97	6.52	8.43	6.51	15.10	40.80	31.78	5.87
Guangdong	8.13	5.19	13.90	38.13	36.68	6.92	8.44	5.70	15.09	42.06	32.14	5.11	8.77	7.04	16.99	43.11	28.81	4.02
Guangxi	8.02	5.18	12.35	39.81	35.35	8.09	8.03	4.57	11.66	41.00	36.90	6.01	7.98	3.29	10.48	45.27	35.40	5.61
Hainan	8.41	5.21	16.88	42.24	29.08	7.35	8.17	5.43	13.57	43.86	28.75	9.50	8.35	5.77	14.60	45.34	26.49	8.65
Chongqing	7.25	3.64	9.80	31.37	44.40	12.28	7.57	4.49	10.93	33.45	42.19	9.70	7.79	4.24	10.68	38.57	39.25	7.80
Sichuan	7.45	3.62	10.52	35.87	39.75	11.53	7.24	4.51	9.20	31.17	43.53	12.56	7.51	4.35	10.05	35.35	40.51	10.24
Guizhou	6.98	4.47	7.89	31.22	41.82	16.98	6.59	2.72	6.38	30.28	44.47	18.79	7.05	3.50	7.45	33.66	42.69	14.58
Yunnan	6.82	3.84	7.46	28.52	45.65	16.37	6.66	3.10	6.76	28.33	46.74	16.50	6.90	3.51	6.76	29.21	48.34	13.29
Tibet	4.40	0.94	2.91	12.03	46.91	44.03	4.16	1.06	2.84	11.22	44.01	45.65	4.71	1.72	3.35	13.21	47.42	37.77
Shanxi	8.26	7.23	15.78	37.94	29.95	10.56	8.30	7.46	15.12	38.79	29.99	9.35	8.51	8.68	15.57	39.25	28.71	8.19
Gansu	7.24	5.67	12.91	29.87	34.87	19.42	6.78	3.30	11.54	30.82	34.83	22.27	7.17	4.49	11.45	32.71	35.49	17.77
Qinghai	6.80	4.51	11.41	28.56	35.68	22.08	6.99	5.95	10.77	27.72	37.55	19.30	7.26	7.47	10.40	26.51	40.43	16.68
Ningxia	7.70	7.16	13.80	31.95	33.75	15.65	7.63	7.27	12.18	33.05	33.77	15.44	8.13	7.65	12.88	37.64	32.84	10.09
Xinjiang	8.49	9.89	13.94	35.22	34.32	7.05	8.30	8.69	11.45	37.51	35.91	6.66	8.56	9.70	11.59	39.37	34.49	4.64

Having a first look of the data, we can get a general understanding of the geographic distribution and development of human capital in China.

Firstly, it is quite obvious that there is a difference in human capital stock among 31 regions. Average years of schooling in Beijing, Tianjin and Shanghai are always in the top rank and the values are much bigger than those of west regions, such as Guizhou, Yunnan and Qinghai. In 1997, average years of schooling in Beijing is 8.04, and that in Shanghai and Tianjin are 7.30 and 6.97 respectively. However, average years of schooling in Qinghai, Yunnan and Guizhou are 4.07, 5.01, and 5.20. Till 2008, average years of schooling in these three western regions have increased to 7.26, 6.90, and 7.05, which are still less than the level of Beijing and Shanghai in 1997.On the other hand, average years of schooling in the three municipalities have already increased to 10.97,10.55, and 9.88 in 2008.

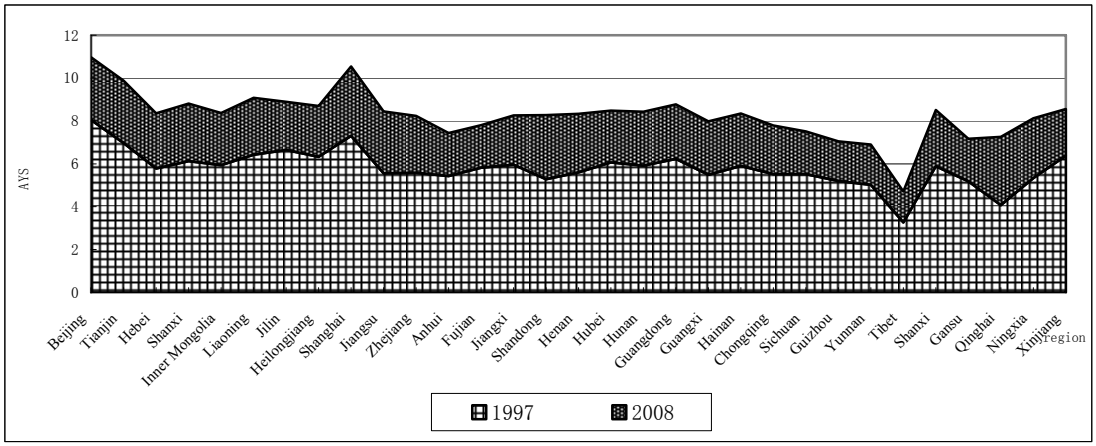

Figure 2.1 Average years of schooling of 31 regions in 1997 and 2008

Secondly, if we look at the structure of human capital, we can find that some regions have an extremely skewed distribution: most residents only have an education level of primary school or middle school; few residents have a college diploma. For example, in 2008, 40.47% of people who aged over six in Qinghai have only an education level of primary school, and people with higher education only accounts for 7.47% in the target population. On the contrary, some regions have a relatively more balanced distribution. Still take Beijing as an example. In 2008, the POC, POH, POM and POP of Beijing are 28.13%, 23.42%, 31.74% and 13.39% respectively, which are not very different from each other. We pick up three regions, Beijing, Hubei (Which is in the central part of China) and Qinghai, and draw the graph based

on data in 2008. It is very clear that they have a quite different structure of human capital.

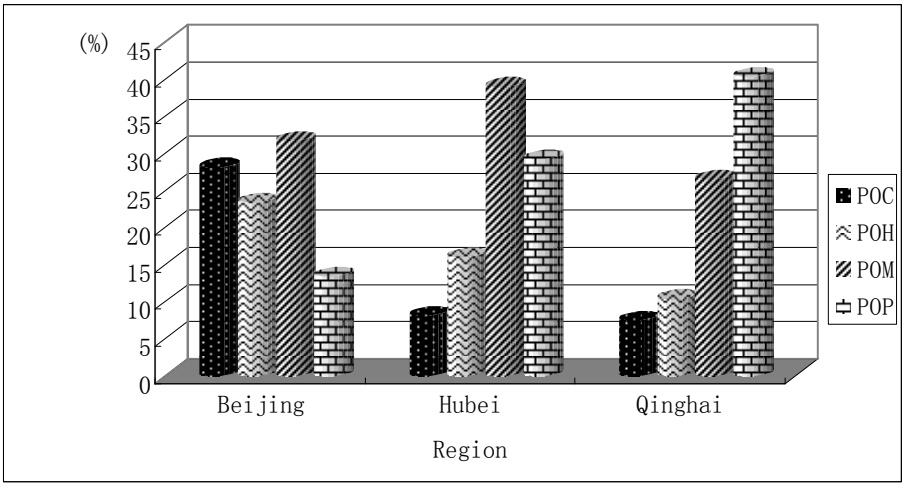

Figure 2.2 Different human capital structures in Beijing, Hubei and Qinghai, 2008

Thirdly, regardless of the differences discussed above, the situation of human capital has improved during the 11 years, and this phenomenon can be observed both in the whole nation and in each single region. Average years of schooling increased. In 1997, the average years of schooling in the whole nation was 7.01, and then increased to 8.26 in 2008. Furthermore, each region also experienced an increase in average years of schooling during the 12 years, and Qinghai has the highest annual increasing rate which is 5.4%. We can take the average of each indicator in 31 regions, and see the fluctuation of them during the 11 years. And to make the interpretation easier, the indicators are scaled such that the average is zero, and divided by the corresponding standard error in that year. It can be obviously seen from the Figure 2.3 that illiteracy rate and people with their highest education level as primary school have declined, while the average of other indicators have increased. If we ignore the changes in population structure, these trends demonstrates that structure of human capital of the whole nation has improved.

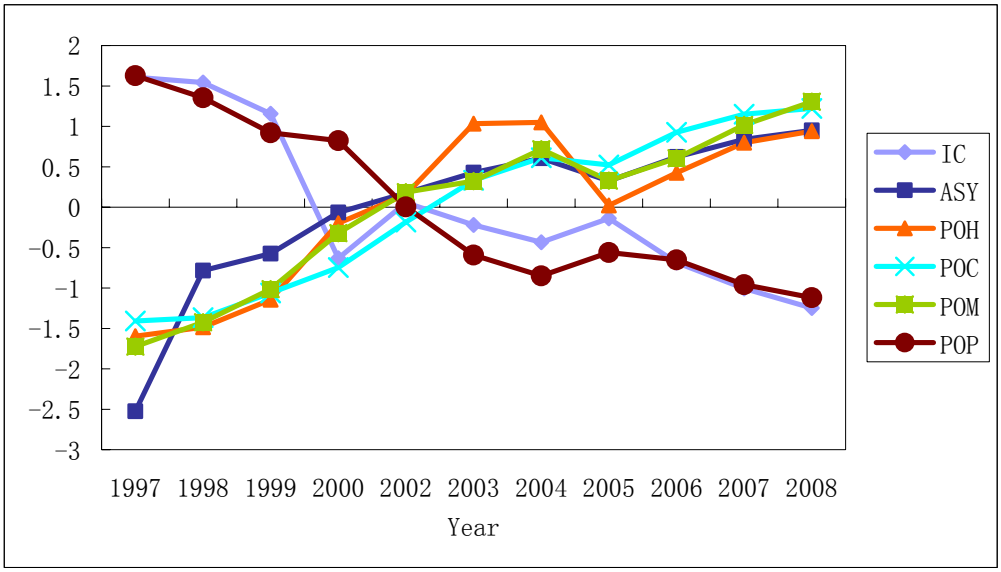

Figure 2.3 Standardized averages of all the 6 indicators from 1997 to 2008

And if we look at each single region, we will find that the structure has been upgraded in some regions, like Beijing and Shanghai, but for some regions, the process of upgrading goes very slowly. We can compare this process between Beijing and Sichuan during the 12 years through the following graphs. It turns out that the structure of human capital in Sichuan does not change much; however, the distribution of human capital in Beijing has graduately shifted to the left, which means a process of upgrading.

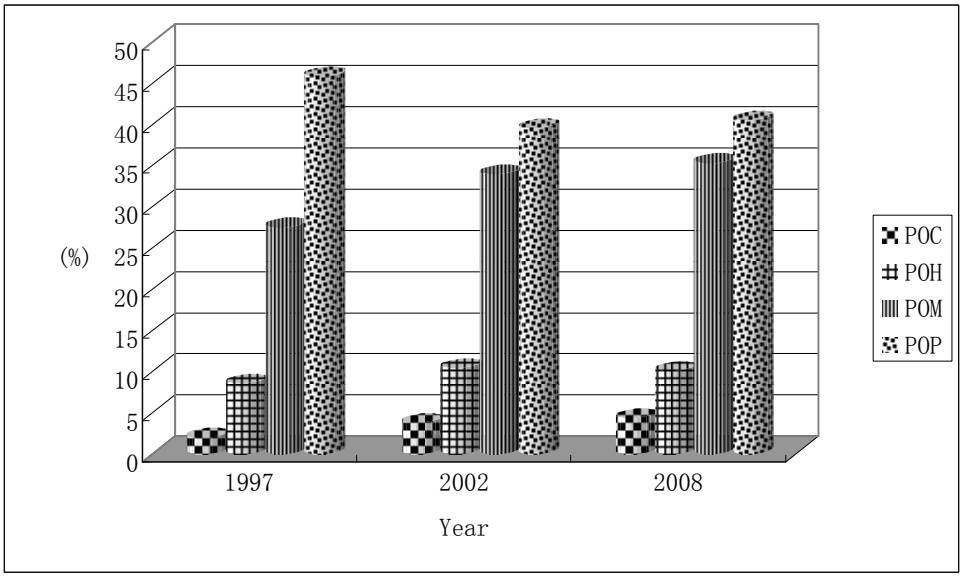

Figure 2.4 the shift in human capital structure of Sichuan

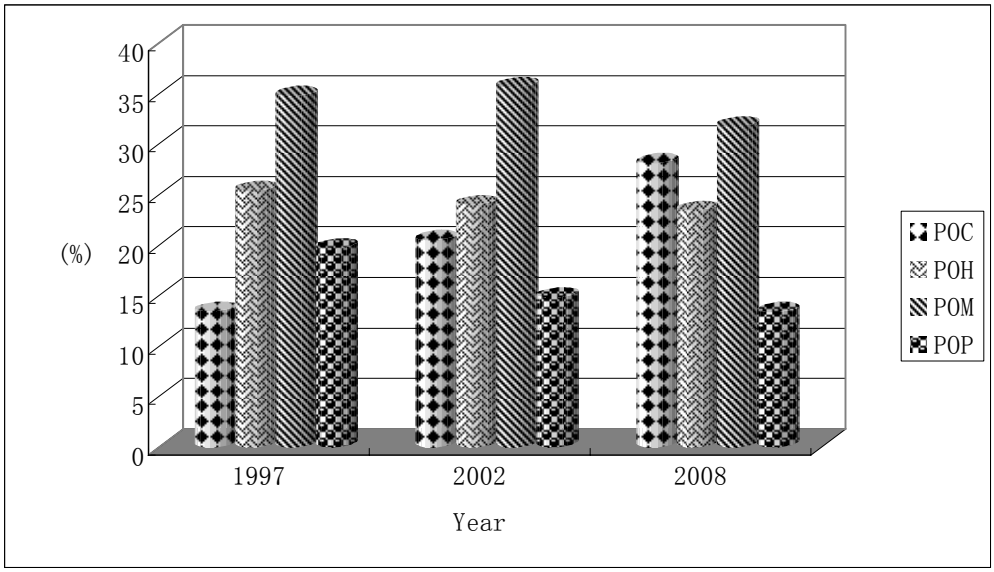

Figure 2.5 The shift in human capital of Beijing

2.3 Map the human capital in China: cluster analysis

We have already had a rough picture of the geographic distribution of human capital in China. However, this first impression is still far from a full understanding. It is of our interests to know which regions are similar in terms of human capital, which differs a lot, and how. Furthermore, there is a common believe in China that human capital in west regions is very poor while the eastern regions did a good job in accumulating human capital. Since there are 31 regions and 6 indicators, it is difficult to make a precise comparison among all 31 regions though the first look of the data, and it is also hard to exam whether the common believe is right. But those are the goal of this section, and to achieve it, we need to take advantage of some econometric tools.

2.3.1 The Method

As we discussed above, we need an econometric method which can help to make a comparison among the 31 regions and find the difference and similarity among them. Cluster analysis enables to meet our needs.

Cluster analysis is a group of algorithms which researchers can use when they want to organize observations in a meaningful structure, that is, to develop taxonomies. According to

a certain distance measure and a certain linkage rule, cluster analysis puts observations in a way that the similarity of observation is maximal if they belong to the same type and minimal if otherwise. There are usually 5 ways of measuring distances: Euclidean distance, Squared Euclidean distance, Manhattan distance, Chebychev distance and Power distance. Among them, Euclidean distance is the most common way, which is simply the geometric distance. Besides the distance measures, each algorithm distinguish itself from the other by linkage rule. Distance measure is only used at the first step when each observation forms a cluster itself; linkage rule defines a way to measure the distance between two clusters which contain more than one observations. Also, there are several linkage rules: single linkage, complete linkage, unweighted pair-group average, weighted pair-group average, and Ward's method etc.

In this study, we choose Euclidean distance as the distance measures, and the formula is as following:

$$d(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

At the same time, we choose Ward's method as linkage rule. This method takes clustering as a variance problem, which distinguishes itself from other method such as single linkage and complete linkage. This method attempts to minimize the error sum of squares (ESS) of any two hypothetical clusters that can be formed at each step. This method tends to have smaller size of clusters, but quite efficient (Wang, 2009).

$$\mathrm{ESS} = \sum_i \sum_j \sum_k |X_{ijk} - \overline{x}_{i\cdot k}|^2,$$

X_{ijk} denote the value for variable k in observation j belonging to cluster i.

In this section, we make use of the latest data, that is to say the data of 2008, and do the cluster analysis. We put all the 6 indicators: AYS, POC, POH, POM, POP，IC) as variables, and try to exam the association of them through Ward's method. One thing worth mention here is that cluster analysis is more like an approach of data mining, thus unlike many other econometric methods, it does not have hypothesis and does not need statistical test. How many clusters one finally chooses depends on the knowledge related to the question on hand and the special needs of the study.

2.3.2 Results

We use SAS 9.0 to accomplish the cluster analysis. The result can be summarized in the dendrogram, which we show below:

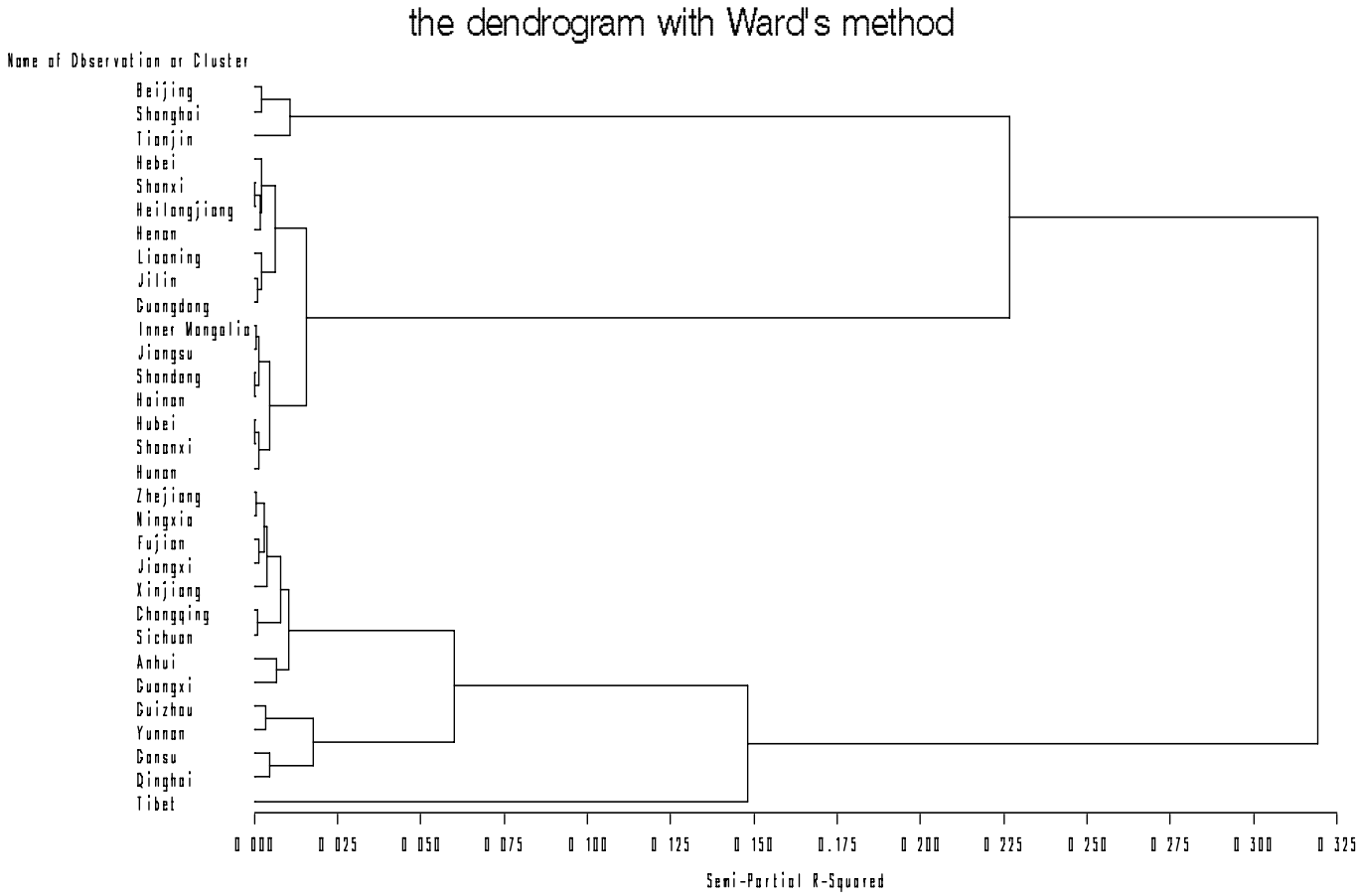

Figure 2.6 The dendrogram with Ward's method: the result of cluster analysis

It can be seen from the dendrogram that 31 regions are categorized to several regions step by step, and at each step the number of clusters which has been formed decreases. At the last step, all regions fall into the same cluster. The X-axis represents the value of Semi-Partial R-Square, from which we can see the similarity of observations in the same cluster: bigger the Semi-Partial R-Square is, less similar they are.

As we have mentioned before, how many numbers of clusters we finally get mainly depends on the purpose of research and relevant knowledge in that field. And our purpose here is to get a clear map of the geographic distribution of human capital in China. If we divided all the regions into too many clusters, it can not give a general picture; if we divide them into too

few clusters, we will lose some important information. For instance, if we categorize them into 3 clusters, it means that Beijing, Shanghai, and Tianjin are in the same cluster, and the rest 28 regions are pooled in another two clusters. In this case, we only know the three municipalities are in the top rank, and we can get few useful conclusions about the other two clusters since each of them includes too many regions to find a common thing in them. Therefore, we finally divide the 31 regions into 5 clusters, which are:

Cluster Ⅰ: Beijing, Tianjin, Shanghai;

Cluster Ⅱ: Hubei, Shaanxi, Shandong, Hainan, Shanxi, Heilongjiang, Inner Mongolia, Jiangsu, Jilin, Guangdong, Hunan, Henan, Hebei, Liaoning;

ClusterⅢ: Zhejiang, Ningxia, Chongqing, Sichuan, Fujian, Jiangxi, Xinjiang, Anhui, Guangxi;

ClusterⅣ: Qinghai, Guizhou, Yunnan, Gansu;

Cluster Ⅴ: Tibet

Furthermore, SAS also reports the statistical characteristics of each cluster, including mean, standard error, minimum value and maximum value. Table 2.3 summarizes these result

Table 2.3 Summary of statistical characteristics of each cluster

Cluster I						Cluster Ⅱ					
Variable	Number of regions	Mean	Std	Min	Max	Variable	Number of regions	Mean	Std	Min	Max
AYS	3	10.46	0.55	9.88	10.97	AYS	14	8.56	0.25	8.28	9.08
POC	3	22.08	6.35	15.46	28.12	POC	14	6.95	1.66	4.72	11.00
POH	3	24.41	0.88	23.42	25.07	POH	14	15.02	1.47	11.99	17.81
POM	3	34.26	2.60	31.74	36.94	POM	14	44.27	3.40	39.00	49.36
POP	3	15.31	2.76	13.39	18.48	POP	14	27.65	1.95	25.27	31.78
IC	3	3.53	0.43	3.11	3.97	IC	14	6.22	1.94	3.48	8.65
Cluster Ⅲ						Cluster Ⅳ					
Variable	Number of regions	Mean	Std	Min	Max	Variable	Number of regions	Mean	Std	Min	Max
AYS	9	7.97	0.37	7.44	8.56	AYS	4	7.09	0.15	6.90	7.26
POC	9	6.12	2.40	3.29	9.70	POC	4	4.74	1.88	3.50	7.47
POH	9	11.82	1.62	10.05	14.90	POH	4	9.02	2.27	6.76	11.45
POM	9	38.11	3.30	34.41	45.27	POM	4	30.53	3.29	26.51	33.66
POP	9	35.67	2.84	32.44	40.51	POP	4	41.74	5.33	35.49	48.34
IC	9	8.79	3.02	4.64	14.49	IC	4	15.58	2.02	13.29	17.77
Cluster Ⅴ											
Variable	Number of regions	Mean	Std	Min	Max						
AYS	1	4.71	.	4.71	4.71						
POC	1	1.72	.	1.72	1.72						
POH	1	3.35	.	3.35	3.35						
POM	1	13.21	.	13.21	13.21						
POP	1	47.42	.	47.42	47.42						
IC	1	37.77	.	37.77	37.77						

If we compare the 5 clusters based on the statistics above, we can find that the average years of schooling decreases from Cluster I to Cluster V: the average years of schooling in Cluster I is 10.46; and that in Cluster Ⅱ, Cluster Ⅲ and Cluster Ⅳ are 8.56, 7.97 and 7.09 respectively; Cluster V, which includes Tibet solely, has a rather low level of average years of schooling, which is 4.71.

The second finding is that there is a "shifting" in the structure of human capital from Cluster I to Cluster V. Comparing to other regiofinalns, a relatively larger percent of residents in regions of Cluster I hold a college diploma: on average, 22.08% of residents in regions of

Cluster I have an college diploma, and this number is almost three times of that of Cluster Ⅱ, four times of that of Cluster Ⅲ, five times of that of Cluster Ⅳ, and twenty times of that of Cluster V. In regions of Cluster Ⅱ and Cluster Ⅲ, the highest education level of most people (44.27%) is middle school, besides the average illiteracy rate in regions of Cluster Ⅲ is 8.79%, which is higher than that in regions of Cluster Ⅱ, which is 6.22%. Most residents in Cluster Ⅳ only attended primary school or middle school; fewer than 14% of residents in those regions have an education level above or equal to high school, and at the same time, illiteracy rate is 15.58%, which is relatively high. Tibet, the only region in Cluster Ⅴ, has an extraordinary high illiteracy rate, which is 37.77%, and less than 19% of the residents have their highest education level above or equal to middle school.

We can use different colors to illustrate the situation of human capital in a region, and the darker the color is, the better the situation of human capital is. By doing this, we finally get a map of geographic distribution of human capital in China[1].

Figure 2.7 The geographic distribution of human capital in China: based on the result of cluster analysis.

As we see from a perspective of geography, Cluster I is consist of 3 municipalities; Cluster Ⅱ is basically composed of Northern regions, some coastal regions and regions in the central part; ClusterⅢ includes some western regions, some regions in the central part, and

[1] The three municipalities are too small to show in a different color, thus they are in the same color of regions in cluster Ⅱ.

two coastal regions (Zhejiang and Fujian); Cluster Ⅳ is mainly consist of western regions; and Cluster Ⅴ includes only Tibet.

This result demonstrated that there is a pattern in the geographic distribution of human capital in China: western regions are mainly less developed than those in the east in terms of human capital; northern regions are more developed than the southern regions; and the abundant human capital in 3 municipalities makes other regions fall far behind.

We have imposed two questions at the beginning of this section, and the last one is whether the common believe is right. By carefully examining the result of cluster analysis, we find that: although in general, economically poor regions are poor in human capital and economically developed regions have relatively abundant human capital, there are indeed some exceptions. Zhejiang and Fujian are both located along the southeast coast line, and are economically developed, but the average years of schooling in these two regions are even less than those in Shaanxi, Shanxi, which are two provinces located in northwest. If we look at the rank of per capita GDP in 2008, Zhejiang is in the fourth place, and Fujian is in the tenth place. However, Shanxi is ranked number 16, and Shaanxi is number 18.

Actually, this is not a new finding. At the beginning of 21st century, some scholars in China has already observed that despite of the success in economic development, Zhejiang falls behind many regions in terms of human capital, and they named it as “Zhejiang Phenomenon”. Many scholars tried to explain the dilemma by reconsidering the relationship between human capital and economic growth and pointed out that at the beginning of transformation, the power of human capital is less obvious than that of infrastructures and etc. However, there may be other reasons. If we think about the data again, we will wonder whether this phenomenon really exists. Just like what many other Chinese researchers have done, we use the average years of schooling among people who aged over six. This group of people is not so closely related to production. It is the human capital in workforce which contributes to economic growth according to growth theories. And a low level of average years of schooling among people who aged over 6 does not necessarily mean a low level of that in workforce. The population structure, the number of students at school and other factors may affect the relationship of the two indicators. So we can not come to the conclusion that human capital contributes little to economic growth in China based on the data among people aged over six.

Even if this phenomenon really exists, there are some other possible reasons which may result in it. For example, culture difference may lead to different attitude in human capital investment. Residents in some regions of Zhejiang and Fujian value "business" more than education, but residents in northern part did the opposite thing. This has already gone beyond our study, but it provides the inspiration of panel data analysis in the following chapters.

Table 2.4 Ranks in per capita GDP and average years of schooling of each region

CLUSTER	REGION	PER CAPITA GDP(in10,000)	RANK OF PER CAPITA GDP	RANK OF AYS
Cluster I	Beijing	6.19	3	1
	Tianjin	6.28	2	3
	Shanghai	7.34	1	2
Cluster II	Hubei	2.05	15	11
	Shaanxi	1.96	18	10
	Shandong	3.30	9	18
	Hainan	1.82	21	16
	Shanxi	2.05	16	6
	Heilongjiang	2.17	14	8
	Inner Mongolia	3.52	8	14
	Jiangsu	4.14	5	12
	Jilin	2.48	11	5
	Guangdong	3.74	6	7
	Hunan	1.81	22	13
	Henan	1.96	17	17
	Hebei	2.32	13	15
	Liaoning	3.72	7	4
ClusterIII	Zhejiang	4.37	4	20
	Ningxia	1.87	20	21
	Chongqing	2.34	12	24
	Sichuan	1.57	24	25
	Fujian	3.17	10	23
	Jiangxi	1.48	27	19
	Xinjiang	1.95	19	9
	Anhui	1.50	26	26
	Guangxi	1.54	25	22
ClusterIV	Qinghai	1.73	23	27
	Guizhou	0.88	31	29
	Yunnan	1.04	30	30
	Gansu	1.21	29	28
Cluster VI	Tibet	1.38	28	31

3 Do the regional differences increase?

The cluster analysis described the regional difference based on one-year's cross-sectional data, thus helped us to have a general view of the recent geographic distribution of human capital. But till now, we do not have any idea about how these differences have evolved with time. To answer this question, we need a new statistics to make good use of the 11-years data we have.

3.1 The Method

To understand the evolution of regional difference in human capital better, we need a statistics which enables us to evaluate regional difference of each indicator of each year. Then we can compare the regional differences among the 6 indicators and also from year to year. If we consider one indicator in a certain year, there are 31 observations which represent the relevant situation in the corresponding region. The regional difference in this indicator is actually the variation of this group of data. A bigger variation means larger regional difference. But by using variation, we can only make the comparison among the values of the same indicator during the 11 years. Which we can not do is the comparison among different indicators. To make this possible, we need a standardized statistics other than variation. That is to say, we need a statistics which could eliminate the difference in scale. Correlated variation just meets the requirement. Chen ,Zhao and Lu ,Ming adopted this method and computed the correlated variation of average years of schooling and other structure indicators, based on data from 1987 to 2001(Chen et al., 2004). Here we will compute more recent values. Another difference from Chen's study is that part of his data is estimated from a panel data model, but mine is all from the reports of SSB in China.

The formula of correlated variation in this paper is as following:

$$CV_{ki} = \frac{\sqrt{\frac{\sum (x_{kij} - \frac{\sum_j x_{kij}}{n})^2}{n-1}}}{\frac{\sum_j x_{kij}}{n}}$$

X_{kij} denotes the value of indication k in region j in year i; CV_{ki} denotes the correlated variation of index k in year i ; n equals to 31.

3.2 The results

We calculate the correlated variation for each indicator during the 11 years, and report the result in the Figure 3.1.

The correlated variation of Average years of schooling in 1997 is 0.14973, and it doesn't change a lot with time. In 2008, this figure decreased very slightly to 0.13481, and the average during the 11 year is 0.148468. We can also see from Figure 3.1 that the regional difference of average years of schooling keeps at a very stable level during the 11 years. On one hand, this reflects the fact that the regional difference of human capital stock is not very big and doesn't change a lot within time. On the other hand, it may also attribute to the other reason which is related to the way of measuring human capital stock: AYS is an indicator with the characteristic of the "Mean" statistics, thus it disregards the structure difference.

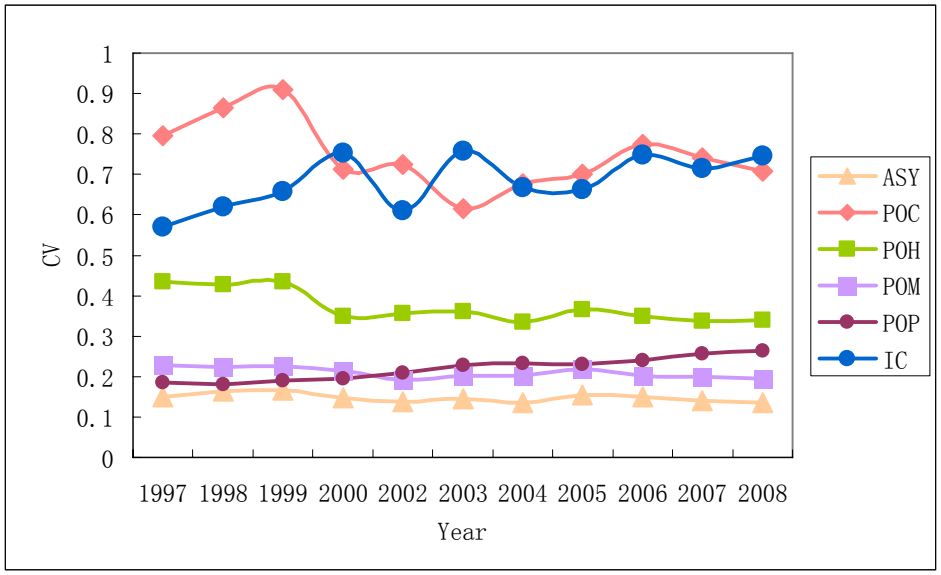

Figure 3.1 Correlated variations of each indicator during 1997-2008

Like average years of schooling, "percentage of middle school" (POM) and "percentage of primary school" (POP) also does not fluctuate greatly. In 1997, the correlated variation of POM is 0.2285, which is a little higher than that of AYS. In most years, it keeps going down,

and decreased to 0.1942 in 2008. The average is 0.2095 during the 11 years. At the same time, although the regional difference of POP in 1997 is smaller than that of POM, it increased graduately: the correlated variation in 1997 is 0.1846, but till 2008, it increased to 0. 2637. And the average is 0.2192 during the 11 years.

Correlated variation of POH in 1997 is 0.4357, which is quite higher than that of POM, POP and AYS. It experienced a relatively great decrease in 2000 and then kept at a very stable level till 2008. In 2008, the figure is 0.3384 and the average is 0.3718.

The rest two indicators, POC and IC carry the biggest regional difference. Besides, the correlated variations of them both fluctuate greatly during the 11 years. The correlated variation of POC is 0.7961 in 1997, and reached its maximum in 1999, which is 0.9101. In 2008, it declined to 0.7075, which still represents a big regional difference.

The correlated variation of illiteracy rate is 0.5692 in 1997, and although it goes up and down, finally it increased to 0.7464 in 2008, even higher than that of POC in the same year. This is a little bit unusual since China has started the campaign of "eliminating illiteracy" for a long time, and also implemented the "Nine-year Compulsory Education" system in 1986, so illiteracy rate in the new generation is supposed to be quite low. Therefore, we recheck the data and find that Tibet has an extremely high illiteracy rate which may affect our result. Then we do the computation again without considering Tibet, and find that this change has little influence on correlated variations of the other 5 indicators, but do make correlated variation of illiteracy rate decrease a lot. On the other hand, this new way of calculation does not change the increasing trend of it.

Table 3.1 Comparison between correlated variation of illiteracy rate with and without Tibet

	1997	1998	1999	2000	2002	2003	2004	2005	2006	2007	2008
CV of IC with Tibet	0.57	0.62	0.66	0.75	0.61	0.76	0.67	0.66	0.75	0.72	0.75
CV of IC without Tibet	0.46	0.48	0.41	0.51	0.46	0.48	0.46	0.50	0.51	0.55	0.51

Now we can make some summaries here. The human capital stock measured in average years of schooling does not differ much from region to region. Nevertheless, the human capital structure differs somehow. The percentage of people with higher education is very different among 31 regions, so do the illiteracy rate. And for POC, the regional difference decreased in

the 11 years, while for IC, it increased. POH also exhibits some degree of regional difference and the other indicators are generally stable and do not exhibit great regional differences.

4 Analysis of the determinants of the unequal distribution

As we discussed in the previous chapters, regional differences in human capital did exists in China during 1997-2008. Indeed, because of the economic transformation staring in 1980s, China exhibits unequal geographic distribution in many aspects. What discussed most is the income gap. However, when it comes to the unequal distribution of human capital, researches either focus on the description of the geographic distribution, or the relationship between human capital and economic growth. Few studies have tried to uncover the veil of the unequal distribution and see what the underlying reasons are. Yang Yang finds that per capita GDP, the education expenditure of centre government and the education expenditure of local governments are the main reasons leading to regional differences of human capital. He used the principal component analysis, which is based on the cross-sectional data in 2008. (Yang, 2009)

Previous analysis and discussions in this paper suggest a mild regional difference in average years of schooling, but a great regional difference in POC. In fact, the unequal distribution of people with higher education is intensively discussed in China and also in other countries, especially the migration of this group of people. In 1974, Bhagwati states that opening-up to migration will lead to “brain drain” in developing countries and harm their economic growth (Bhagwati, 1974). However, Stark, Helmenstein and Prskawetz proved that under certain conditions, the opening-up could have a “brain gain” effect (Stark et.al, 1998). Discussions as such are all about the international talent migration, not the migration within one country. At the same time, although Chinese scholars have noticed the internal talent migration for a long time, there are not many empirical studies, neither do theoretical models.

Thus, in this chapter, we will make good use of the data on hand, and try to find the determinants of the unequal distribution of average years of schooling, as well as the percentage of people with higher education (POC). In the final section of this chapter, we will give a brief discuss of the validity of the estimation. And a theoretical model of internal migration will be derived in the next chapter.

4.1 What determines geographic distribution of human capital stock?

4.1.1 The method

Here we want to find what the key determinants of geographic distribution of human capital are. And the dependent variable will be average years of schooling of each region in each year. Because we have data of 31 regions, from 1997 to 2008(except 2001), it is very nature to think about the panel data regression. What is more, the advantages of panel data regression enable to reduce the omitted variable bias to a limited level in our study.

Panel data regression is good at controlling two types of omitted variable bias: one is the bias caused by the omitted variables which vary from entity to entity, but do not change with time; the other is the bias caused by those which are constant across entities but evolve over time. In our study, there do exist some factors which may affect the average years of schooling, and also POC, and which are hard to measure and then not possible to be put into the econometric model. For example, people's attitude towards education has changed during the 11 years, and taking education has been accepted graduately by more and more people. This change has the same influence on all regions across the country, and is supposed to increase the average years of schooling and POC. This belongs to "time fixed effect". Meanwhile, due to different historical reasons, different regions may value education differently, which is called the "entity fixed effects" .Take Fujian province and Liaoning province as examples. Fujian has a history of going aboard for small business, and most of the first –generation-migrants who smuggled themselves abroad came from Fujian. So taking education is not a very important thing in their point of view. While in Liaoning province, and most other northern regions of China, parents value children's education a lot, and this consequently leads to a higher human capital stock. Other factors which may affect the average years of schooling and POC are the location and nature environment of a region. Northern part of China is extremely cold and the air has been heavily polluted due to the long-period development of heavy industrial base. It should have a negative effect on human capital stock. This is only one case of how regions' special climate and nature environment influent its accumulation of human capital. In a word, it is usually difficult to measure such factors, and omitting these variables could lead to bias in

corresponding estimators. However, the fixed effect model of panel data provides a possible approach to eliminate the bias resulted from the impossibility of measuring.

Therefore, the econometric model can be written as following:

$$AYS_{it} = \beta \bullet X_{it} + \alpha_i + \lambda_t + u_{it} \qquad (4.1)$$

β and X are two vectors: $X = (x_1, x_2, ...x_n)$ is the vector of independent variables; $\beta = (\beta_1, \beta_2 ... \beta_n)$ is the vector of coefficients corresponding to different independent variables. α_i is the entity (region) fixed effects, and λ_t is the time fixed effect. Where i denotes region i, and t denotes year t. μ_{it} is the residual.

According to Greene, for a single variable x, the estimator is unbiased and consistent when the following assumptions are met (Greene, 2008):

1. $E(ASY_{it} \mid x_{it}, \alpha_i, \lambda_t) = \beta_0 + \beta \bullet x_{it} + \alpha_i + \lambda_t$;

2. No perfect multicollinearity;

3. $Var\ E(\mu it \mid x_{it}, \alpha_i, \lambda_t) = 0$;

4. $Cov(\mu_{it}, u_{st}) = \begin{cases} \sigma^2 I, \text{when i=s} \\ 0 \quad \text{otherwise} \end{cases}$

5. $Cov(\mu_{it}, u_{id}) = 0$ where $\forall t \neq d$

6. $\mu \sim N(0, \sigma^2 I)$

The first assumption is the linearity assumption, which assumes that the relationship between x and y is linear. The third assumption says that the regressors are exogenous, and if this assumption is violated, the estimator is biased. The fourth assumption states that variance with the same region in the same year is constant, otherwise is zero. The fifth one is about the autocorrelation, and it assumes that there is no correlation in error terms across different years even within the same region. Among the six assumptions, the violation of assumption 3 often arises in multiple regressions, due to the omitted variable problem. As a result, we will pay special attention to it when selecting proper independent variables.

4.1.2 The Variables

Now we have figured out the proper approach to find out the determinants of average years of schooling, it is the time to discuss what the possible determinants are, in another words, what

the independent variables should be in this regression. And Generally speaking, we will base our selection of regressors on the study of two fields: human capital investment theories and brain drain theories, and to see whether those variables emphasized in theoretical models work in the same way in China.

Because whether to take education is a decision of investment according to many scholars in this field, the elements which have an influence on individual's decision may also affect average years of schooling in a region. The underlying logic is that human capital stock results from all its residents' decision of education investment: if more people choose to take education or to increase their education level, the average years of schooling in a region must rise up. At the same time, educated people could choose freely where to work, thus the elements which are evolved in the migration decision can also affect the human capital stock in a region. When a region enables to attract more educated people, the human capital stock in this region must increase. On the other hand, even if a region has an abundant resource of students, but fails in the competition for talents in job markets, it may finally turn out to be poor in terms of human capital stock.

Income is the first regressor coming into mind. Becker, one of the founders of human capital theory pointed out in his famous paper in 1964 that increasing the rate of return in education works as an incentive of investing in education. And the rate of return in education is positively related to the net earning, which is determined to a large extent by the income (Becker, 1964). Besides, according to Harris-Todaro migration theory, it is the expected income that influences the potential migrants' decision (Ray, 1998). In many recent brain drain studies, like Miyagiwa's study of scale economy in education and brain drain(Miyagiwa, 1991), Vidal's study of the spillover effect of emigration on human capital formation(Vidal,1998), expected income plays a fundamental role in people's migration decision, which then affects the human capital accumulation in a certain region. Therefore, we include income as one of the regressors in our model with the hypothesis that regions with higher income are richer in terms of human capital stock. To be more specific, we take the average disposable income of urban citizens in a region as one regressor.

The unemployment rate is a major determinant of people's real cost associated with investment of education, and plays a role in people's education decision (Blöndal et al., 2002).

A higher employment rate of educated people could encourage people to take education, and as a result, could cause an increase in human capital stock. Besides, the probability of successful migration can lead to a "brain drain" or a "brain gain", depending on how big the probability is (Chen, 2008). In our scenario, we argue that the probability of successful migration is positively related to the employment rate in the destination region .Because in modern China, there are few obstacles in internal migration, and it can be seen as a free choice. In order to settle down in a certain region, the only thing one needs to do is finding a job there to survive. Therefore, we will include unemployment rate as an independent variable in the regression. In fact, unemployment rate among educated people is a better indicator, but the data is only available in years when there is a national census.

Cost is the other variable worth while considering from the perspective of human capital investment and migration. Cost in a general sense affects the human capital stock of a certain region through two mechanisms. The first mechanism is through tuition and fees, which is the individual's private cost of taking education. And keeping other factors constant, like earnings, public financial support, a higher tuition and miscellaneous fees work as disincentives of taking education, thus it should be negatively related to human capital stock. However, in practical, the data of education cost is not available for all regions. As a result, we are not able to include it in our model. The validity problem caused by this exclusion will be discussed at the end of this chapter. The second mechanism is through living cost, which influences the accumulation of human capital stock in a region by affecting people's migration decision. Although living cost is seldom integrated into either the human capital investment models or the migration models, it doesn't mean living cost do not affect the human capital stock of a region at all. High living expenditure in big cities in China has caused some graduates return back home, which increases the human capital in small cities and decreases that in big cities. Thus we use the average annual consumptional expenditure of urban citizens in each region to indicate the living cost, and add it into the regression model.

Besides income and cost which are usually evolved in people's cost-benefits analysis of education, there are other nonpecuniary benefits which are also under people's consideration. Wilson developed a structural model in which utility-maximizing individuals choose their education level in response to economic return of education as well as neighborhood

characteristics. According to Wilson, neighborhood characteristics include location, the composition of residents, and economic status (Wilson, 2001). If we take the possibility of migration into account, those characteristics also influence individual's choice of migration destination. According to the Push-pull Theory, the early migration theory, convenient life and good medical treatment system are all examples of pull factors. Thus, we include 3 independent variables to reflect the neighborhood characteristics, or in other words, nonpecuniary benefits: the ratio of urban residents to rural residents; the number of street lights per each city; and the number of health personnel per 10, 000 inhabitants. The first variable is closely related to the urbanization process in a certain region. More urban residents mean a more open culture, but also a more crowding living environment in cities. While the first effect may provide incentive for taking education and attract more educated people, the second effect may drive educated people out of the region. Therefore, the impact of this variable on the dependent variable may go to either direction. The number of street lights per cities is usually is an indicator of development, and is considered able to reflect the development of infrastructure in a region What is more, it plays an important role in adding beauty to a region at night, and also affects people's perception about the security situation in that region. On one hand, in areas with more street lights, fewer crimes will happen; on the other hand, more street lights could help drivers to have a better view and then lead to a reduction in the number of car accidents. Based on these reasons, people may prefer being educated and living in a region with more street lights. Thus, number of street lights per cities is assumed to be positively related to both average years of schooling and POC. Health personnel in China include 3 types of people: certified doctors, assistant doctors, and registered nurses. This last variable represents the level of medical treatment in a region, which plays an important role in individual's migration decision: people always want to have better access to medical treatment. Thus, this regressor is assumed to positively related to the two dependent variables.

Government's education expenditure is anther variable which is often discussed by Chinese scholars when it comes to the study of human capital. In some researches, it is selected as an indicator of human capital (Lu et.al, 2010); and in some researchers, government's education expenditure is considered be able to evaluate a region's ability of cultivating talents, and as a

result, has significant influence on human capital accumulation of a region (Gu, 2006). In fact, governments' expenditure, or subsidy on education is anther source of human capital investment besides private investment, and it has crucial impact on human capital stock through improving school quality, according to Wilson's study based on American data (Wilson, 2001), and at the same time, it is considered to be able to encourage people attending school in China, especially those in rural areas who are not able to afford education alone (Heckman, 2004). The discussions above have offered solid theoretical foundation for us to include government's education expenditure in the regression. To avoid being disturbing by different population size, we finally take the government's education expenditure per person as a regressor.

Based on the discussion above, we can modify the regression model as following:

$$AYS_{it} = \beta_0 + \beta_1 \bullet income + \beta_2 \bullet unem\_rate + \beta_3 \bullet cost + \beta_4 \bullet rur + \beta_5 \bullet streetlight + \beta_6 per\_hp + \beta_7 \bullet per\_e\_expenditure + \alpha_i + \lambda_t + \mu_{it} \quad (4.2)$$

Although we have selected 7 independent variables, it does not mean omitted variable bias does not exist in this regression .As long as there are omitted variables which are correlated with the included regressors and at the same time are a determinant of the dependent variable, omitted variable bias will arise (Stock and Watson, 2007). Thus, it is worth thinking twice before we proceed further with the data.

We should first notice that since there are already 7 variables, each of them may be biased due to the violation of assumption 3. However, we should also notice that the regressor itself not only be a potential determinant, but also can be a control variable for other regressors. For example, income is positively related with the number of street lights per cities, because usually an economically developed region has more street lights and its residents have a higher level of average income. The similar relationship can be also found between the number of street lights per cities and the number of health personnel per 10,000 inhabitants, because they are both positively dependent on how developed a region is. As a result, the combination of such variables enables to reduce the omitted variable bias. Then, the next question is, are there any omitted variables which are not included in our framework and thus not added in model (4.2)?

The first possible variable coming into mind is the number of students. It is obvious that number of students has an important effect on average years of schooling: holding other things equal, a higher number of students leads to a higher level of average years of schooling. Meanwhile, number of students is associated with income, which represents the economical status of a region. That is, in regions where people are richer, people's education decisions are not limited by the budget constrain, and as a result more people could go to school. Based on the discussions above, we need to add number of students into the model. To eliminating the effect of population size, we finally add the student ratio in the whole population into the model:

$$AYS_{it} = \beta_0 + \beta_1 \bullet income + \beta_2 \bullet unem\_rate + \beta_3 \bullet cost + \beta_4 \bullet rur + \beta_5 \bullet streetlight + \beta_6 per\_hp + \beta_7 \bullet per\_e\_expenditure + \beta_8 \bullet student\_ratio + \alpha_i + \lambda_t + \mu_{it} \quad (4.3)$$

Nevertheless, (4.3) is still not the final model we will use, because there are anther variable which we have neglected. The average years of schooling in our study is based on the population aged over six, so if many people are in the school age and fewer people are the labor force, the average years of schooling may be smaller. In other words, population structure will influence the level of human capital stock. Furthermore, population structure is also correlated with the student ratio, and also the average annual income as well as expenditures, since different age groups have different earning abilities and consumptional habits. Consequently, if we do not include population structure as a regressor, the omitted bias problem will arise. To measure the population structure, we use the ratio of people over 15, and modify the model as:

$$AYS_{it} = \beta_0 + \beta_1 \bullet income + \beta_2 \bullet unem\_rate + \beta_3 \bullet cost + \beta_4 \bullet rur + \beta_5 \bullet streetlight + \beta_6 per\_hp + \beta_7 \bullet per\_e\_expenditure + \beta_8 \bullet student\_ratio + \beta_9 \bullet rp + \alpha_i + \lambda_t + \mu_{it} \quad (4.4)$$

Till now, we accomplished the econometric model we will use, and the independent variables are summarized in Table 4.1.

Table 4.1 Summary of the independent variables in regression to average years of schooling

Variable	Name	Unit
Average annual disposable income of urban residents	income	10,000RMB
Average annual consumptional expenditure of urban residents	cost	10,000RMB
Unemployment rate	unem_rate	%
Population ratio of urban residents to rural residents	rur	%
The number of street lights per city	streetlight	one
The number of health personnel per 10,000 inhabitants	per_hp	person
Government's education expenditure per student	per_e_education	RMB/person
Student ratio	Student_ratio	person
Ratio of people aged over 15	rp	%

4.1.3 Data Source

In this section, we will give a brief introduction of the data which is going to be used in the regression.

The main data source is the "China's Statistical Yearbook" from 1998 to 2009. We can obtain "Average disposable income of urban residents", "Average consumptional expenditure of urban residents" directly from the yearbook in the corresponding year. To compute "the number of street lights per city", we use the data "number of street lights" in this yearbook, and divide it by "the number of cities", which is also available in the yearbook. For some years, the yearbooks report "the number of health personnel per 10,000 inhabitants", but for most years, they only report "the number of health personnel", so we have to divide it by population provided by the same yearbook. To compute "Per capita government's education expenditure", we collected the data of "government's education expenditure" from the yearbook, and divided it by population. "The China's statistic yearbook" reports the number of students every year, and then we divide it by population to get the "student ratio". "Ratio of people aged over 15" is simply based on the information of "Age Composition by Region"

provided by the yearbook.

The other two variables, “Unemployment rate” and “Population ratio of urban residents to rural residents” are from different sources. “Unemployment rate” is “the registered unemployment rate in urban areas”, which is found at “China Labor Statistic Yearbook”; and “Population ratio of urban residents to rural residents” is from “The Information Website of Development Research Center of the State Council”.

4.1.4 The results

We use the statistical software SAS.9.0 to implement the regression. In practical, we use 10 regional binary indicators and 30 time binary indicators to represent the entity fixed effects and time fixed effects, respectively.

The estimated values of coefficients in equation are summarized in table 4.2. The dependent variable is average years of schooling. In the first column, we use the multiple regression model, which does not include entity fixed effect indicators as well as time fixed effect indicators. The regression equation of this the model is vey similar to (4.2), but without α_i and λ_t. Results of equation (4.2) are summarized in column 2, and we use the fix effects model, adding both entity fixed effect indicators and time fixed effect indicators. In column 3, we added “student ratio” to control for the omitted variable bias, and the corresponding regression model is (4.3). The last column represents the results of equation (4.4), in which we add “population ratio of people aged over 15” to control the omitted variable bias. The degree of freedom for this model is 49 since we have 40 binary variables, 7 explanatory variables and 2 control variables.

Table 4.2 The panel data regression analysis of determinants of average years of schooling

Dependent variable: average years of schooling (years per person)				
Regressor	(1)	(2)	(3)	(4)
income	-0.038	0.06**	0.06**	0.082**
	(0.056)	(0.02)	(0.02)	(0.019)
cost	1.572**	-0.71**	-0.71**	-0.901***
	(0.237)	(0.23)	(0.23)	(0.215)
unem_rate	-0.083	0.012	0.012	0.002
	(0.054)	(0.023)	(0.023)	(0.02)
rur	0.056**	0.005	0.005	-0.003
	(0.007)	(0.005)	(0.005)	(0.005)
streetlight	-0.000001	0.000003**	0.000003**	0.000004**
	(0.000001)	(<0.000001)	(<0.000001)	(<0.000001)
per_hp	-0.009	0.004	0.004	0.0099**
	(0.007)	(0.004)	(0.004)	(0.004)
per_e_education	0.000006	0.00006*	0.00006*	0.00007**
	(0.00005)	(0.00003)	(0.00003)	(0.00002)
Student_ratio	0.013	----	0.001	0.026*
	(0.017)		(0.011)	(0.012)
rp	----	----	----	0.083**
				(0.012)
entity effects?	no	yes	yes	yes
time effects?	no	yes	yes	yes
observations	331	331	331	331
Adjusted-R2	0.6066	0.9721	0.9720	0.9758

The data used include 31 regions, from 1997 to 2008 and except 2001. The region and time fixed effects are included in (2), (3) and (4), but not reported. Standard errors are given in parentheses under the coefficients. The individual coefficient are statistically significant at the *5%level, or **1%level.

Comparing the results in column2 with that in column 1, we find that firstly the Adjusted-R^2 has increased from 0.61 to 0.97. And the coefficients which are statistically significant in the two specifications are different. In specification (1), only cost and the population ratio of urban residents to rural residents are statistically significant. However, in specification (2), income, cost, the number of street lights per cities and government's education expenditure per capita are all statistically significant. Furthermore, the signs before the coefficients of the first three variables have changed, comparing to specification (1). Although the result of

Specification (3) does not differ much from that of the specification (2), the combination of student ratio and population ratio of people who aged over 15 does improved the model, as column (4) shows. The coefficient of the added variable is statistically significant; the coefficients of the number of health personnel per 10,000 inhabitants and student ratio become statistically significant after the modification; Adjusted-R^2 increased. Thus, regression (4.4) is the most reliable regression and we will focus on it in the following discussion[1]. And we will discuss the validity of this regression later.

Income has a positive effect on the average years of schooling, holding other variables constant. To be more specific, controlling for other variables, the average years of schooling will increase by 0.082 years if the average annual disposable income of urban residents increases by 10,000RMB. This effect is not as great as that of the cost. A decrease in the average annual consumptional expenditure by 10,000RMB could lead to a rise of average years of schooling by 0.901. Thus, the income difference between two regions could cause the difference of human capital stock, so does the cost difference. Furthermore, the regional cost difference has a greater effect on the regional difference of human capital stock.

The number of street lights per cities, number of health personnel per 10,000 inhabitants, and the population ratio of urban residents to rural residents are the three indicators we selected to represent nonpecuniary benefits, or the neighborhood characteristics. The coefficients before the first two variables are both statistically significant at the 1% level. However, the third one does not have statistically significant association on average years of schooling in our model, even at a 10% significance level. Holding others constant, a region owning 10,000 more street lights per cities tends to be 0.04 years higher in average years of schooling. And a region where every 10,000 inhabitants owned 10 more health personnel tends to be 0.09 years higher in average years of schooling.

The result in our study supports other Chinese scholars' conclusion that government's education expenditure has an influence on the human capital stock in that region (Gu, 2006; Yang, 2009). Controlling for other variables, a local government which spends 10,000RMB more in education for each of its resident, gains a level of average years of schooling which is

[1] We actually also did the log-linear regression afterwards, however, the result of specification (4.4) is still preferred since it meets our gut feeling better.

0.7 years higher than that of other regions. Nevertheless, this effect is smaller comparing to that of decreasing cost but bigger than that of increasing income.

To our surprise, unemployment rate is not statistically significant even at a 10% significance level. Based on our previous discussion, other factors are given, if the probability of finding a job in certain region is bigger than other regions, it will attract more educated people. The surprising result here may be because we have used the registered unemployment rate of all people rather than the unemployment rate among educated people. Another possible reason is difference in the situation of job markets really does not affect people's education decision and migration decision, due to the asymmetric information or lack of sensitivity. That is to say, people do not have full information about the true situation of job market, or even they have, their responses to it are slow and slight.

The student ratio also has a statistically significant effect on average years of schooling at a 5% significance level, although it is not the variable which we want to exam through empirical study. A 1% increase in student ratio render the average years of schooling to increase by 0.026 years, holding other variables constant.

One thing worth mentioning here is that the coefficient of student ratio is not statistically significant before we add the second control variable. Thus the second indicator helps us to decrease the omitted variable bias. And the indicator itself also has a statistically significant effect on average years of schooling. Just as we have discussed, if this indicator is very small, a population is too young to have a high level of human capital stock. Our results suggest that a 1% increase in this indicator leads to a 0.083 years increase in the average years of schooling. Of course, we are not very interested in the effect of the population structure on human capital stock, comparing to the effects of other indicators, and we select this indicator only on the purpose of controlling omitted variable bias.

To conclude, most indicators we selected are estimated to have the same influence on the geographic distribution of human capital stock as theories suggested or are the same with other empirical studies. At the same time, some are not, for example, the unemployment rate. In the next section, we will discuss another variable, the percentage of people with higher education, which exhibits the biggest regional difference, and to see what the determinants are and also compare the results with what we have found in this section.

4.2 What determines geographic distribution of talents?

We define talents as people with higher education, and it is worth while discussing the geographic distribution of talents in a separate section because of two reasons. Firstly, we have shown that, the correlated variation of POC is the highest among the 6 indicators through the 11 years, and the average is around 0.75. Those facts suggest a big regional difference in talent stock. Secondly, some scholars have argued that a key reason resulting in this unequal distribution is the migration of talents. Yang argued that in regions located at the central part or the western part, like Anhui, Henan, Sichuan, Guizhou and Gansu, the fruits of high education have been "eaten up" by the migration for jobs (Yang, 2009). Since migration of people with higher education is relatively easier and thus more frequent than people with a lower level of education, the mechanism which leads to the geographic distribution of talents is more complicated. Therefore, we will continue our econometric analysis in this section, as what we have done to the average years of schooling.

4.2.1 The methods and variables

The econometric methods and the framework on which the selection of variables is based are similar to the analysis of average years of schooling. We will use the entity and time fixed effect model to implement the estimation, with a panel data which include 31 regions and 11years (from 1997 to 2008 and except 2001). The dependent variable is the percentage of people with a college diploma or above, i.e. POC. The basic model can be written as:

$$POC_{it} = \beta' \bullet X'_{it} + \alpha'_i + \lambda'_t + u_{it} \qquad (4.4)$$

Where β' is a vector of coefficents and X' is a vector of regressors. α'_i is the entity effects and λ'_t is the time effects. μ_{it} is the residual. i denotes region i and t denotes year t.

Except for some small adjustments, the variables we select here are as the same as what we added in the model (4.3). Average annual disposable income of urban residents, average annual consumptional expenditure of urban residents, unemployment rate and the three nonpecuniary indicators (population ratio of urban residents to rural residents; the number of

street lights per city; the number of health personnel per 10,000 inhabitants) are kept as independent variables of POC.

Nevertheless, instead of government's education expenditure per capita, here we take government's expenditure of high education per college students as the independent variable, which is more closely related on the dependent variable.

The specification of our model can be expressed as following:

$$POC_{it} = \beta'_0 + \beta'_1 \bullet income + \beta'_2 \bullet unem\_rate + \beta'_3 \bullet cost + \beta'_4 \bullet rur + \beta'_5 \bullet streetlight + \beta'_6 per\_hp + \beta'_7 \bullet per\_he\_expenditure + \alpha'_i + \lambda'_t + \mu_{it} \quad (4.5)$$

And to control the omitted variable bias, we add the college student ratio of the whole population in the model, which is similar to the approach we used in the regression against average years of schooling.

$$POC_{it} = \beta'_0 + \beta'_1 \bullet income + \beta'_2 \bullet unem\_rate + \beta'_3 \bullet cost + \beta'_4 \bullet rur + \beta'_5 \bullet streetlight + \beta'_6 per\_hp + \beta'_7 \bullet per\_he\_expenditure + \beta'_8 \bullet cstudentratio + \alpha'_i + \lambda'_t + \mu_{it} \quad (4.6)$$

Furthermore, we add population ratio of people who aged over 15 to control for additional omitted variable bias, and the mechanism is the same as we discussed in 4.1.2.Based on discussions above, we write the final model of POC as following:

$$POC_{it} = \beta'_0 + \beta'_1 \bullet income + \beta'_2 \bullet unem\_rate + \beta'_3 \bullet cost + \beta'_4 \bullet rur + \beta'_5 \bullet streetlight + \beta'_6 per\_hp + \beta'_7 \bullet per\_he\_expenditure + \beta'_8 \bullet cstudentratio + \beta'_9 \bullet rp + \alpha'_i + \lambda'_t + \mu_{it} \quad (4.7)$$

Table 4.3 Summary of independent variables in regression to POC

Variable	Name	Unit
Average annual disposable income of urban residents	income	10,000RMB
Average annual consumptional expenditure of urban residents	cost	10,000RMB
Unemployment rate	unem_rate	%
Population ratio of urban residents to rural residents	rur	%
The number of street lights per city	streetlight	one
The number of health personnel per 10,000 inhabitants	per_hp	person
Government's expenditure of high education per college student	per_e_education	10,000RMB
Ratio of college students	ncstudentratio	%
Ratio of people aged over 15	rp	%

Variables in light gray are those which slightly differ from Table 4.1.

4.2.2 Data Source

The data source of the first six variables has been introduced in section 4.1.3. Here we only illustrate the data source of the other two variables.

Government's expenditure of high education per college students is from the "China Educational Finance Statistical Yearbook", however, it dose not state how it is computed. And the number of college students is the "Number of students in Undergraduate or Specialized Courses in Institutions of Higher Education". Thus, postgraduates and various doctors are not included in the calculation.

4.2.3 The Results

Again, four regressions are implemented and the results are summarized in Table 4.4. The first column gives the result of the regression similar to (4.5), but without the entity and time fixed effects. Then we added 30 indicators of entity and 10 indicators of year, and the result is in column 2. Adjusted-R^2 has improved, from 0.85 to 0.95 after the adjustment in regression model and average annual consumptional expenditure becomes statistically insignificant. Although a further modification according to (4.6) does not improve the Adjusted-R^2 to a large extent, the result in column 3 suggests that college student ratio has a very big impact on POC. Furthermore, the coefficient of the government's expenditure in higher education per college students, which is statistically significant in the first two regressions, is not statistically significant in specification (3). Finally, we add the population ratio of people aged over 15 according to model (4.7), and the result in column (4) shows that it does have a statistically significant impact on POC, and because of this modification, the coefficient of college student ratio becomes smaller, but the coefficient of income becomes bigger. To conclude, the last regression model is the most reliable one[1], therefore, we will base our discussion on it. Similarly, the degree of freedom for the model is also 49.

[1] Here we also did the log-linear regression, and again, the result does not meet our gut feeling well, thus we will continue with what we already had.

Table 4.4 The panel data regression analysis of determinants of POC

Dependent variable: POC（%）				
Regressor	(1)	(2)	(3)	(4)
income	0.551**	0.931**	0.832**	0.867**
	(0.125)	(0.095)	(0.096)	(0.096)
cost	4.723**	0.129	-0.274	-0.572
	(0.579)	(1.107)	(1.083)	(1.082)
unem_rate	0.173	-0.026	-0.114	-0.114
	(0.118)	(0.018)	(0.108)	(0.108)
rur	0.018	-0.01	-0.029	-0.044
	(0.015)	(0.026)	(0.025)	(0.025)
streetlight	0.00003**	0.00005**	0.00005**	0.00005**
	(0.000003)	(0.000004)	(0.000004)	(0.000004)
per_hp	0.117	0.041*	0.06**	0.07**
	(0.015)	(0.018)	(0.02)	(0.02)
per_he_education	-1.048**	0.231*	0.323	0.327
	(0.312)	(0.369)	(0.360)	(0.357)
cstudentratio	----	----	1.423**	1.301**
			(0.357)	(0.358)
rp	----	----	----	0.141*
				(0.059)
entity effects?	no	yes	yes	yes
time effects?	no	yes	yes	yes
observations	331	331	331	331
Adjusted-R2	0.8483	0.9521	0.9545	0.9552

The data used include 31 regions, from 1997 to 2008 and except 2001. The region and time fixed effects are included in (2), (3) and (4), but not reported. Standard errors are given in parentheses under the coefficients. The individual coefficient are statistically significant at the *5%level, or **1%level.

There are five variables which have statistically significant effect on POC, and two of them are added to the model to control omitted variable bias. Average annual disposable income has a positive impact on POC: controlling for other variables: if the average annual disposable income increases by 10,000 RMB, the percentage of people holding a college diploma or above will increase by 0.867%. Therefore, regional income difference does not only lead to the regional difference of human capital stock, but also the regional difference of talent stock.

Two of the three nonpecuniary indictors are statistically significant, just as the situation in regression against average years of schooling. The coefficient of the number of street lights per city is 0.00005, which means that other things equal, the region which owns 10,000 more street lights per cities tends to be 0.5% higher in POC. The number of health personnel per 10,000 inhabitants is also statistically significant at a 10% significance level, and the effect is relatively great. An increase of number of health personnel per 10,000 inhabitants by 10 will lead the POC to increase by 0.7%, holding other variables constant. Notice that these two indicators are also statistically significant in model (4.3), thus, like the effect of regional income difference, the regional differences in modernization as well as medical treatment can cause regional differences in both human capital stock and talent stock.

Now it is of our interests to have a look at those indicators which have a statistically significant influence on average years of education, but not on POC. Average annual consumptional expenditure and government's expenditure in education are two such indicators.

As we discussed in 4.1.2, although few scholars has taken living cost into account when constructing structural model of brain drain, it may has an impact on talent stock of a region. However, the result in our study does not support our original hypothesis, and also seems deviating from the reality. Just as we mentioned before, the high expenditure in Beijing, Shanghai and Guangdong has forced some employees return to their home towns, and this trend is intensively discussed in China as "escaping from big cities" these days. So what are the reasons behind the conflict between our result and the reality? Recall that annual average consumptional cost is a statistically significant indicator in the regression against average years of education. The average years of schooling as an indicator of human capital stock, results from the actions of people with various education backgrounds, but POC is only affected by the action of talents. Furthermore, although some people choose to leave big cities, most of them choose to stay. Therefore, it is reasonable to explain the phenomenon from a perspective of different behavior patterns among people. One possible reason is that for most people with a good education background, the high living cost in the short run does not matter much, and they emphasize more on the income stream in the coming future and the development in their career. And the result of this behavior pattern is that living cost is

statistically insignificant in the empirical study.

Unemployment rate is not statistically significant, and we have given the possible explanation in 4.1.4, thus there is not necessary to repeat it again.

Let's turn our eyes upon government's expenditure in higher education. Although the relationship between government's expenditure and average years of schooling has been studied by many scholars, there are few papers about the relationship between government's expenditure in higher education and the talent stock. Our study here suggests that there is not a statistically significant association between the two variables, and it is a surprising result to some extent. One possible reason is that: government's subsidy for primary school or middle school releases the budget constraints for some poor families and makes the corresponding education free. The direct consequence is more people can attend school and the average years of schooling increases. However, despite of the subsidy from government, tuitions and miscellaneous fees for higher education are still very high, so the effect of releasing budget constraint is not obvious and efficient here. Thus, government's expenditure in higher education may only improve the quality of students, but does not increase the quantity.

College student ratio has a very closely positive association with POC, which is not surprising. According to Table 4.4, a 1% increase in college student ratio could induce POC to increase by 1.3%. However, this relationship is more mathematical rather than having any economic meaning. So does the relationship between POC and population structure. On the other hand, studies of some scholars demonstrated that some regions having abundant number of college students suffer from a brain drain: the flowing-out of talents (Yang, 2009).

We have already discussed the determinants of the geographic distribution of human capital stock as well as the talent stock. But stock is a static concept, and in fact, the formation of this static concept can be partly attributed to a dynamic concept: migration of educated people. In fact, we have had some brief discussion about it when selecting independent variables of the regression. Internal Migration of educated people is very important especially when we are talking about the regional difference of talent stock. So far we are not able to illustrate this process in China precisely and also not able to obtain some interesting results. Furthermore, we have not discussed the determinants of geographic distribution in

human capital in labor force. Since an empirical study is impossible due to the unavailability of data, a theoretical model is necessary and feasible. Therefore, in Chapter 5, we will try to construct an internal migration model, to complement our discussion about geographic distribution of human capital. But what comes in the next section, is an examination of the validity of our econometric study.

4.3 Validity of the Estimation

We have already found the key determinants for average years of schooling and the POC, respectively. And the results show that: income, cost, government's education expenditure per student, the number of street lights per cities and the number of health personnel per 10,000 inhabitants have statistically significant impact on average years of schooling in a region; income, the number of street lights per cities and the number of health personnel per 10,000 inhabitants are statistically significant associated with POC. Although our major purpose is to figure out which are the main determinants, and we are not very interested in the magnitude of the impact, it is still worth to discuss about the validity of the econometric study, because some problems may cause such a serious bias that we may get the wrong answer to our main question.

4.3.1 Omitted Variable Bias

We have paid special attention to the omitted variable bias problem while selecting the proper regression model. Two things have been done to control for the omitted variable bias. The first thing is that we use a panel data regression, which includes entity and time fixed effects and thus enables us to control for some unobserved variables; the second thing is we added another two additional control variables: the student ratio(the college student ratio in regression against POC), and the population ratio of people who aged over 15. By doing these, omitted variable bias can be thought as well controlled and declined to a limited level. However, as we mentioned before, we do not have information of education cost, so it is impossible to include this in the two regressions. If it is related with any existing independent variables, omitted variable bias will arise.

4.3.2 Simultaneous Causality

Regressions we did in the previous sections helped to find out the variables which have statistically significant association with human capital stock as well as talent stock. And till now, we assume that this "association" is the causality running from independent variables to dependent variables. In fact, it may be the case that a change in dependent variable also causes the changes in independent variables. And we call it the "simultaneous causality". And in such circumstances, the estimators provided by OLS regression are biased and inconsistent (Stock and Watson, 2007). It is time to have a second look at the independent variables we have selected, and see whether a possibility of simultaneous causality exists. Considering that student ratio (college student ratio in (4.7)) and population ratio of people aged over 15 are two control variables, which does not have economic implication, we will exclude them from the following discussion.

A lower unemployment rate in certain region may enhance the incentive of taking education and migration, thus is supposed to increase human capital stock and talent stock in the corresponding region. At the same time, gathering of educated people may have two opposite effects on unemployment rate. On one hand, the abundant resource of educated people can contribute to the economic growth in the region, and more jobs are created, which pull down the unemployment rate. On the other hand, job competition could be fiercer comparing to the regions which own fewer educated people, and as a result, the unemployment rate tends to rise up. Regardless of which effect dominants, the existence of simultaneous causality causes unemployment rate to be correlated with the error term, and consequently, the estimator is biased and inconsistent.

Simultaneous causality may also arise between the number of health personnel per 10,000 inhabitants and human capital stock as well as talent stock. Because health personnel must meet certain criterion, it is possible that a larger number of educated people, especially a larger number of people with higher education, could lead to a lager number of health personnel per 10,000 inhabitants. Similarly, in this case, a simultaneous causality bias arises and the estimator is inconsistent.

One way to solve the problems addressed above is using the instrumental variables regression.

We will not continue the study here, but a further discussion and modification of our study is appreciated.

4.3.3 Errors-in-variables

Errors–in-variables bias arises when an independent variable is not measured precisely. And there are several sources of this type of bias. The data we used are from surveys in the corresponding year, and except for 2000, surveys in the other years are sample survey. Furthermore, the way of selecting the sample has changed from 1997 to 2008. For example, the information about income and cost are from the "Sample Survey on Urban Household". And before 2002, the objects of urban household are non-farm household, which is determined by "Hukou", but after 2002, they are changed to households in the district areas of all city and county towns. Thus, there may be errors-in-variables due to the change of sampling, and also due to responders' imprecise answer to questions. The other possible source is that registered unemployment rate is an inaccurate measure of the situation in job market for educated people. As we discussed before, unemployment of educated people is a better indicator in (4.4), and unemployment of people with higher education background is a better indicator in (4.7). However, the data is not available in most years.

Through the discussion above, we know that errors-in-variables may exist in our study, however, to a large extent it is due to the limitation of data availability, which is hard to improve in a short time. Once again, modification and improvement of our study is appreciated, but we will go no further in this paper.

5 A model of internal migration

5.1 Background

We have studied the determinants of geographic distribution of human capital stock in the whole population. However, due to the unavailability of data, we are not able to examine the elements which affect the geographic distribution of human capital in labor force[1]. As we have discussed, migration of educated people or talents, is a dynamic process, and which could affect the formation of human capital stock in the whole population, especially that in the labor force. Because the number of educated labor force in a region can be considered solely being determined after the two migration processes. Thus, if we can construct a theoretical model to explain the mechanism behind the internal migration, it will complement our study in the geographic distribution of human capital in labor force, and also will help us to understand the internal migration better.

Previous papers and publications study the international migration of educated people from the perspective of "brain drain". If a large number of skilled people have migrated to foreign countries, the source country is considered experiencing a "brain drain". By saying brain drain, scholars mean loss of educated labor force. Brain drain is usually regarded as an economic cost: on one hand, the source country loses many skilled people, who are able to contribute to the economic growth in the source country according to the endogenous growth theory; secondly, there is also out-flow of government's subsidy in education with the out-low of skilled people. Earlier researches in brain drain found that the immigration of skilled people has negative implications in the source countries' welfare (Bhagwati, 1974). And besides losing number of people, source country also loses human capital in immigration (Katz et.al, 1987). However, in 1990s, scholars became focus on anther possible result brought by immigration of skilled people, and the problem of international migration has been discussed intensively during this period. Under a model of scale economy, it is demonstrated that it is those professionals possessing intermediate-level abilities who are hurt by brain drain,

[1] We do not have information about the education attainment in people aged over 15, thus it is impossible to calculate the average years of schooling and POC.

but at the same time, brain drain could raise the education and income levels of a host country (Miyagiwa, 1991). If migration is not a certainty, a brain drain may increase average productivity and equality in the source economy (Mountford, 1997). And the level of human capital formation in the source country can be positively correlated with the probability of emigration; an incidentally surge in emigration can lead the source country out of an under-development trap (Vidal, 1998). The condition under which a strictly positive probability of employment in a foreign country raises the level of human capital of the source country has been specified (Stark et al., 1998). It is also demonstrated that under some conditions, a "brain effects" will dominant "the drain effect" (Beine, 2001). In these papers, the probability of migration is assumed exogenous in the context of international migration, because there is a visa control in foreign countries. Attempts to endogenize it have been made. And it is found that if probability of migration based upon the "threshold effects" of average human capital, and if households perceive that there is a high probability of migration in the future, they will invest more in their education, thereby increasing the accumulation of human capital, which will in turn induce a higher probability of migration (Chen, 2008).

The migration trend of educated people among countries can be also observed within a country. It is shows that some provinces experienced a gain in human beings but a net loss in human capital, or vice versa; and some provinces are more adversely affected by the flows of human capital than others (Fan, 2007). However, comparing with international migration of educated people, there are not many theoretical models aiming at explaining internal migration of educated people. It is the rural-urban migration which is mostly discussed when it comes to the internal migration in developing countries like China. In fact, internal migration of educated people is also a well-noticed problem in China recently, and it is considered to have hurt the poor regions and benefits the rich.

So, a natural question next is: is the existing brain drain model suitable for the situation in China? In the settings of those models, residents in poor country, who decided to take education, are educated in their home country. This "home country effect" can be resulted from two facts: on one hand, the opportunity of studying abroad is rare, and residents in poor country neither have much freedom to choose where to be educated, nor could afford the expenditure in foreign countries. On the other hand, the visas offered by foreign country are

very limited, and to increase the competitiveness in obtaining a visa, residents in source country have to first finish higher education at home. Nevertheless, when it comes to the study of brain drain within a country, this framework seems not reasonable any more. At least for the case in China, youths can go to universities in whatever region they like after graduation from high school. The only requirement is passing the entrance examination, which is basically a just criterion. In this case, two steps have been evolved in the migration process: migration for higher education and migration for job. The two trends together decide the talent ratio in the whole population, and the talent ratio in labor force, not solely the second process. So it is of our interests to build a model in such scenario, and include all the potential determinants we have discussed in empirical study, and finally see how they affect the talent stock in a region.

5.2 The Model

In this model, we would like to incorporate the potential determinants we have discussed in the empirical studies, which are also based on previous researches. These parameters are: income, education cost, living cost, unemployment rate, non-pecuniary benefits.

Assume that there is a group of people, and each of them possesses different level of latent ability. e_i denotes the latent ability of individual i. We follow Mountford's assumption that these latent abilities is distributed in the interval [0, E], and the corresponding density function is g (e_i) (Mountford, 1997). The group of people is going to decide where to take higher education, and there are two regions they can choose from: region A and region B. The education costs in the two regions are C_A and C_B, respectively. Region A is more economically developed, so it can offer a higher starting wage W_A for graduates; region B offers a starting wage W_B, and $W_A > W_B > 0$.

We further assume that the probability of getting a job in region B is 1, regardless of where you graduated from. It means that as long as you want to find a job in region B, you can make it. The underlying reason is that region B is lack of skilled labor[1]. And it can result from many facts. For example, bad weather, low level of welfare, closed culture and out-dated industrial

[1] By saying skilled labor, here we mean people who have a college diploma or above.

structure, etc. On the contrary, the competition for job in region A is very fierce: for people who graduate from region A, the percentage of people who succeed in finding a job there is π_A ; for people who graduate from region B, the percentage of people who could finally get a job there is π_B. π_A, π_B are exogenous and are determined by the economic environment in region A. $\pi_A > \pi_B$ because employers in region A prefer graduates in local universities since they have better information about them. Meanwhile, different individuals have different expectations on the probability of getting a job in region A, which are dependent on the levels of latent ability they possess and also where they have been educated. For people educated in region A, the probability function is α (e_i); for people educated in region B, the probability function is $\beta(e_i)$, and they have the following properties:

(1) $\alpha(e_i), \beta(e_i) \in [0,1]$; $\alpha'(e_i) > 0, \beta'(e_i) > 0$; $\alpha(e_i) > \beta(e_i)$ for all $e_i \in [0, E]$.

(2) $\alpha(e_i) - \beta(e_i)$ is a decreasing function of e_i

Property (1) implies that α (e_i) and $\beta(e_i)$ are both strictly increasing function of e_i, which means if one possesses a higher level of latent ability, she will think it is easier for her to get a job in region A. Since individuals form their expectations on the basis of information they have, they also take account of employers' preference for local graduates, and that is why we have the last inequality of property (1). However, when the latent ability one possesses is higher, the difference in probabilities caused by different education places becomes smaller, that is what property (2) implies.

There are two decisions to make for the group of people, and we will analyze them one by one.

Decision Ⅰ: Where to be educated?

The group of people can be divided into two subgroups according to where they want to work, and it is determined by their preferences which we will discuss later. For this moment, let's call people who want to work in region A "A-fans", and those who want to work in region B "B-fans". They compare the net benefit they can obtain from the two regions, and choose where to be educated.

For "A-fans", they will choose to be educated in region A as long as:

$$\alpha(e_i)\bullet W_A + (1-\alpha(e_i))\bullet W_B - C_A\cdot(1+r) \geq \beta(e_i)\bullet W_A + (1-\beta(e_i))\bullet W_B - C_B\cdot(1+r) \qquad 5.2.1$$

Where r is the interest rate, and is fixed.

The left hand side represents the expected net benefits of being educated in region A. The sum of the first two items is the expected income, and the third item is the education cost. Similarly, the right hand side is the expected net benefits of being educated in region B.

For "B-fans", they will choose to be educated in region A as long as:

$$W_B - C_A\cdot(1+r) \geq W_B - C_B\cdot(1+r) \qquad 5.2.2$$

Notice that the probability of getting a job in region B is 1, and since they are "B-fans", expected income of them is just W_B.

If $C_A = C_B$, it is obvious that "B-fans" will be indifferent about where to be educated, and "A-fans" will choose to be educated in region A, because:

$\alpha(ei) > \beta(ei)$ for all e_i and $W_A > W_B \Rightarrow \alpha(e_i)\bullet W_A + (1-\alpha(e_i))\bullet W_B > \beta(e_i)\bullet W_A + (1-\beta(e_i))\bullet W_B$

Now let us further assume $C_A > C_B$. And this is usually the case because it is more expensive to attend colleges in more developed regions.

For "B-fans", they now prefer to be educated in region B. What about "A-fans"? They will choose to be educated in region A so long as inequality 5. 2. 1 is satisfied.

$$\alpha(e_i)\bullet W_A + (1-\alpha(e_i))\bullet W_B - C_A\cdot(1+r) > \beta(e_i)\bullet W_A + (1-\beta(e_i))\bullet W_B - C_B\cdot(1+r)$$

$$\Leftrightarrow \qquad \alpha(e_i) - \beta(e_i) > \frac{(C_A - C_B)(1+r)}{W_A - W_B}$$

Let $\varphi(e_i) = \alpha(e_i) - \beta(e_i)$, so $\varphi'(e_i) < 0$. [1]

The right hand side is positive, and if $\frac{(C_A - C_B)(1+r)}{W_A - W_B} < 1$, there exists an e**[2], where:

$$\varphi(e^{**}) = \frac{(C_A - C_B)(1+r)}{W_A - W_B} \qquad 5.2.3$$

"A-fans" with $e_i \geq e^{**}$ will choose to be educated in region B; and the others will choose to be educated in region A. The threshold level of latent ability is:

---

[1] Since $\varphi(e_i) = \alpha(e_i) - \beta(e_i)$, $\varphi'(e_i) = \alpha'(e_i) - \beta'(e_i) < 0$, according to our assumption

[2] When $\frac{(C_A - C_B)(1+r)}{W_A - W_B} < 1$, the right hand side is belong to the interval (0,1), which is also the range of $\varphi(e_i)$ .According to the Intermediate value theorem, there must exist an e** which satisfies 5.2.3.

$$e^{**} = \varphi^{-1}\left(\frac{(C_A - C_B)(1+r)}{W_A - W_B}\right) \qquad 5.2.4$$

Decision Ⅱ: Where to work?

As we mentioned above, the group of people can be categorized by "A-fans" and "B-fans", but we have not discussed how one type distinguishes itself from the other type. In the following discussion, we will show how this self-distinguish is realized.

Assume people's utility function is a transformation of Cobb-Douglass utility function, which is writt as following:

$$u(l, x) = a\ln l + b\ln x + c\ln y$$

l is leisure, and x is comsumption goods with price p. y represents non-pecuniary benefits (including access to medical treatment, appearance of the city,etc). a, b and c are different weights which individuals assign to different elements, and they are all positive. Notice that y is a given value, which is determined by the development of the corresponding region, and individuals only choose among different vaules of l and x.

Let $\bar{L}$ denotes the maximum hours a worker can work, W denotes the wage level,which is a constant for a certain region. Then the agent's maximization problem is as following:

$$\underset{x,l}{Max}\ \ a\ln l + b\ln x + c\ln y$$

$$s.t\quad w(\bar{L} - l) = px$$

It is easy to derive the corresponding indirect utility function, which is:

$$v(w, p, \bar{L}, y) = a\ln(a\bar{L}) + b\ln\left(\frac{bw\bar{L}}{p}\right) + c\ln y - (a+b)\ln(a+b)$$ [1]

Now we assume that for different individuals, c is different, and c is an increasing and continuous function of $e_i, c'(e_i) > 0$ and $c(e_i) \in (0, +\infty)$. The explanation of this assumption is that individuals with higher latent ability put a greater weight on the non-pecuniary benefits.

People will prefer to work in region A as long as:

$$v_A(w_A, p_A, \bar{L}, y_A) \geq v_B(w_B, p_B, \bar{L}, y_B) \qquad 5.2.5$$

where $v_A(\cdot)$and $v_B(\cdot)$ are the indirect utility functions people will have if working in region A and region B, respectively.

To solve 5.2.5, we have

[1] See appendix 1.

$$5.2.5 \quad \Leftrightarrow \quad a\ln(a\bar{L}) + b\bullet\ln(\frac{bw_A l}{p_A}) + c(e_i)\ln y_A \geq a\ln(a\bar{L}) + b\bullet\ln(\frac{bw_B l}{p_B}) + c(e_i)\ln y_B$$

$$\Leftrightarrow \qquad c(e_i)\bullet\ln(\frac{y_A}{y_B}) \geq b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B}) \qquad 5.2.6$$

Since region A is much more developed than region B, $y_A > y_B$, so $\ln(\frac{y_A}{y_B}) > 0$. Therefore, there are two cases in the solution of 5.2.6, and we will discuss it one by one.

Case (1): $\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B}) \leqslant 0 \Leftrightarrow \frac{W_B}{P_B} \leq \frac{W_A}{P_A}$

In this case, the left hand of 5.2.6 is strictly positive, and the right hand of 5.2.6 is negative or zero. Thus, 5.2.6 can be satisfied for any e_i. That is to say, nobody wants to work in region B; all of them want to work in region A. Notice that $\frac{w}{p}$ is actually the real wage in a region, and the interpretation of case (1) is the real wage in region B is smaller than that of region A.

Case (2): $\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B}) < 0 \Leftrightarrow \frac{W_B}{P_B} > \frac{W_A}{P_A}$

$$5.2.6 \quad \Leftrightarrow c(e_i) \geq \frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)}$$

It can be proved that if $c(E) \geq \frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)}$, there exists an e*, which makes $c(e^*) = \frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)}$ [1]. And if $c(E) < \frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)}$, 5.2.6 will not hold regardless of e_i.

In the later case, nobody wants to work in region A. Because this is an extreme situation, which does not coincide with the reality in China, thus, here we focus on the first situation.

When $e_i \geqslant e^*$, 5.2.6 can be satisfied. So "A-fans" are those with latent ability $e_i \geqslant e^*$, otherwise, they are "B-fans". The threshold level of ability in this case is:

[1] See appendix 2.

$$e^* = C^{-1}\left(\frac{b\ln\left(\frac{W_B \bullet P_A}{W_A \bullet P_B}\right)}{(\ln y_A - \ln y_B)}\right)$$

5.3 The effects on number of educated labor force

We will focus on region B in this section, and see how the changes in parameters affect the number of educated labor force.

Let us combine the results we have obtained in decision Ⅰ and decision Ⅱ, there are actually 3 scenarios

(1) $\frac{W_B}{P_B} \le \frac{W_A}{P_A}$;

(2) $\frac{W_B}{P_B} > \frac{W_A}{P_A}$ and $e^* \ge e^{**}$;

(3) $\frac{W_B}{P_B} > \frac{W_A}{P_A}$ and $e^* < e^{**}$.

However, the first two scenarios sound not realistic. In the first scenario, everybody wants to work in region A because 5.2.6 holds for all e_i; in the second scenario, everybody chooses to take education in region B, because all "A-fans" possess a level of latent ability higher than e**, which enables them to enjoy the lower education cost in region B. Thus, we only focus on the case which is more similar to the reality in China, and that is the last case. In this intermediate case, both regions can be a choice of people, either as place to work or to be educated. And we further assume that any change in any parameter will not fall into a range where this scenario does not hold any more.

It is necessary to summarize the situation in the intermediate scenario here to prepare for further discussion. In this scenario, people with $e_i \ge e^*$ want to work in region A, but some of them, with $e_i \in (e^{**}, E]$, choose to take education in region B; the others choose to be educated in region A. On the other hand, People with $e_i < e^*$, want to work in region B, and also choose to be educated in region B.

Because of the fierce competition in region A, not all people who want to work there enable to find a job. People who failed at finding a job in region A go to work in region B, then contribute to the human capital there. As a result, the number of educated labor force in region

B can be expressed as following:

$$N(w_A, w_B, p_A, p_B, y_A, y_B, C_A, C_B) = \int_0^{e^*} g(e_i)de_i + (1-\pi_A)\bullet\int_{e^*}^{e^{**}} g(e_i)de_i + (1-\pi_B)\bullet\int_{e^{**}}^{E} g(e_i)de_i$$

5.3.1

where $(1-\pi_A)$ and $(1-\pi_B)$ are two kinds of unemployment rates in region A: $(1-\pi_A)$ is for people who is graduated from region A; $(1-\pi_B)$ is for people who graduated from region B. And $\pi_A > \pi_B$ as we have discussed.

The first term of 5.3.1 is the number of "B-fans", and the second term is "A-fans" who are educated in region A and have to work in region B; the third term is the number of "A-fans" educated in region B and also failed to find a job in region A. In fact, those "A-fans" now become a source of human capital in region B. It is clearly that any change in any parameters will lead to a change in $N(\cdot)$, which is a change in the number of educated labor force.

To see how exactly the parameters affect the number of educated labor force in region B, we can calculate the partial derivatives of $N(\cdot)$, and a positive partial derivative means that increasing the corresponding parameter leads to gain in educated labor force in region B; a negative partial derivative means that decreasing the corresponding parameter leads to a human capital gain in labor force.

We summarize the results in Table 5.1, and the proofs of results are given in appendix 3.

Table 5.1 Summary of the results of partial derivatives

Partial derivatives	Sign	If increase the parameter
$\frac{\partial N}{\partial y_B}, \frac{\partial N}{\partial C_A}, \frac{\partial N}{\partial P_A}, , \frac{\partial N}{\partial W_B}$	positive	Gain in human capital
$\frac{\partial N}{\partial y_A}, \frac{\partial N}{\partial C_B}, \frac{\partial N}{\partial P_B}, \frac{\partial N}{\partial W_A}$	negative	Loss in human capital

To conclude, according to the setting of this model, if there is an increase in the education cost (C_A) and living cost (P_A) in region A, the number of educated labor force in region B will increase. At the same time, if the non-pecuniary benefits (y_A) in region A, like the culture, the

access to entrainment and etc increase, the number of educated labor force in region B will decline. And if the starting wage offered by region A goes up, it will attract more educated people, and region B will suffer a loss of educated labor force. The same change in the corresponding parameters in region B will lead to an opposite result.

5.4 A brief discussion about the preferences

We assumed that people with different level of latent ability assign different weight to non-pecuniary benefits, which will influence their choices of working place. And if we look at 5.2.6 carefully, we will find that the weight assigned to leisure actually doesn't matter. It is the combination of b and c which matters. In fact, when we tried to explain why cost is not a statistical significant variable in regression 4.7, we have already mentioned the possible existence of different preferences. Therefore, in this section, we will have a further discussion about modeling the preference. Notice that we still focus on the intermediate case here.

Instead of $c(e_i)$, we assume that the ratio between c and b is a function of e_i: $\frac{c}{b} = \alpha(e_i)$.

And $\alpha(e_i) > 0,\ \alpha'(e_i) > 0$. Therefore, to solve the inequity 5.2.5, we have:

$$c \bullet \ln(\frac{y_A}{y_B}) \geq b \ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})$$

$$\Leftrightarrow \quad \frac{c}{b} \geq \frac{\ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{\ln(\frac{y_A}{y_B})}$$

$$\Leftrightarrow \quad \alpha(e_i) \geq \frac{\ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{\ln(\frac{y_A}{y_B})} \qquad 5.4.1$$

Consequently, when people assign different weights to consumption goods and non-pecuniary benefits, they may choose different regions to work in. And 5.4.1 suggests that only when the ratio between the two weights exceeds $\frac{\ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{\ln(\frac{y_A}{y_B})}$, there are people who want to work in region A. Otherwise, people choose to work in region B, despite of the higher wage, more non-pecuniary benefits offered by region A.

Now let's jump out of the frame work we have build in 5.2, and just consider decision Ⅱ. Furthermore, we release the assumption that α is a function of e_i. Instead, we assume α is a function of many factors, that is to say, $\alpha = \alpha(\bar{x})$. And $\bar{x}$ is a vector which includes latent ability as well as some other elements. Thus, according to this way of modeling preference, any element in $\bar{x}$ could affect the number of educated people in a region through affecting people's preference and then people's decision about where to work. Possible elements include family background, social network, etc. However, we are not able to exam these impacts unless we have access to data in individual level.

The best way to obtain the related individual data set may be conducting a sample survey with designed questionnaire. We can first ask for information about education background as well as other basic information which is supposed to have an association with the formation of preferences. For example gender, major in college, what kind of jobs parents take and etc. And then we can include regional characteristics in the questionnaire and ask responders how they value all this characters. Finally a regression could help to reveal the preferences of people with different education background, and also what factors have influenced the formation of preferences. With these results, we can make better-targeted and more effective policies to improve human capital stock and upgrade human capital structure.

Because dataset described above is not available at this moment, we could not do the empirical study about preferences here. However, more effort can be made in this field, and of course should be made in the future.

5.5 Remarks

Discussions and results above provide some inspirations to local governments about how to enhance human capital development in a region. According to the results above, increasing starting wage for graduates, reducing education cost as well as living cost, and improving nonpecuniary benefits such as available public goods could all increase human capital stock in a region. Governments may not enable to force various organizations to increase starting wage, but they can use tax as a tool to adjust disposable income of individuals. However, in

China, each region has the same income tax system: the same threshold and the same tax rate. If local governments could have more freedom in adjusting income tax, less developed regions may have another chance to improve their situation of human capital. Comparing to the first approach, the last two approaches may be more feasible for local governments. There are many ways to reduce education cost and living cost: reduce tuitions, increase education subsidies, control inflation, provide low-rent apartments and etc. Meanwhile, providing public goods is a responsibility of government: decrease the incidences of crime, improve public transportation and maintain a diversity culture could all enhance the attractiveness of a region. Nevertheless, all these potential policies can lead to a surge in government expenditure, thus a more efficient allocation of government revenue should be achieved.

Of course, as we have mentioned, this theoretical model is aimed to analyze human capital embedded in workforce, which we do not have relevant data either. If we want to exam the results of our theoretical model in the real world, we could use the similar method in Chapter 4. However, first we need to calculate the average years of schooling in workforce and obtain a panel data set. Indeed, this task is of great importance in China since no one has done it. Average years of schooling used in most papers are referring to that in people aged over six. As we discussed before, to study the relationship between human capital and economic growth, we need human capital in workforce, not in the whole population. Thereafter, a progress in this field will not only help to draw a map of human capital in workforce in China, and also will lead to new findings in other related fields.

6 Conclusions and suggestions

In this paper, we collected data during 1997-2008 of 31 regions, and built a panel data set. Based on it, we studied the geographic distribution of human capital in China, described its status quo, analyzed how regional difference in human capital changes with time, and found out the main determinants of the unequal distribution. The main conclusions can be summarized as following:

(1) There is a regional difference in human capital stock and human capital structure. Generally speaking, northern regions are better than southern regions, and eastern as well as central regions are better than western regions. However, not all eastern regions, which are more economically developed, also have an advantage in human capital; Zhejiang and Shanxi are good examples. Besides, percentage of people with higher education exhibits the highest regional difference.

(2) Each region differs to some extent in average years of schooling, but the difference is not very large and keeps at a stable level during 1997-2008. Regions differ most in percentage of people with higher education, then that of high school education and illiteracy rate; regions do not differ much in percentage of people with middle school education and primary school education, which are included in the "Nine year compulsory education system". For most of the structure indicators, the regional difference tends to decline, and only the regional difference of illiteracy rate and percentage of people with primary school education have increased.

(3) Income, government's education expenditure and living cost have statistically significant influence on average years of schooling. The affect is positive for the first two, and negative for living cost. However, only income has statistically significant influence on percentage of people with higher education, the rest two does not matter. And two nonpecuniary benefits indicators: the number of street lights per cities and the number of health personnel per 10,000 inhabitants have statistically significant impacts on the two dependent variables, and the impacts are positive. Unemployment rate is not statistically significant in the two regressions.

(4) According to our model of internal migration, education cost and living cost in a certain region are negatively related with human capital stock in labor force in the same region; the effects of starting wage and nonpecuniary benefits are in the opposite way.

There are some interesting and meaningful finds, which provide some inspiration in making relative policies:

First, government's expenditure in higher education seems not correlated with the percentage of people with higher education. Although compulsory education and higher education both have positive external effects, the first one is made free by government while the second one is not. Thus, despite of government's subsidy for higher education, individuals still have to pay for it by themselves. Thus, government's expenditure may improve the quality of higher education, but does not encourage many poor people to take education. A possible solution for this is to build a complete credit system for higher education, for example, make sure that most students have access to low-interest education loan. By doing so, government's expenditure may play a more significant role in increasing the number of people with higher education.

Secondly, disposable income has strong effect on both average years of schooling and percentage of people with higher education. Although income is dependent on the employers, disposable income is actually under the control of government. An appropriate reduction on tax could release individuals' budget constrain, and make taking education affordable for more people.

Thirdly, construct a better information transmission system of job market. Unemployment rate do not affect neither of the two dependent variables in our study. Despite of the possible errors-in-variables, lack of full information may be a main cause. Full information could guide people's behavior better, and will help to redistribute human capital in a more efficient way, which then could balance the development in human capital among regions

Finally, high living cost seems not able to drive people with higher education out of big cities. On the other hand, non-pecuniary benefits could attract more talents and also increase the average years of schooling. As a result, for regions which aim at attracting more people with higher education, direct subsidies in living cost may not very effective; improvements

nonpecuniary benefits such as medical treatment level or the appearance of city may do a better job.

Reference

Becker, G. S. (1964): “Human capital: A theoretical and empirical analysis, with special reference to education”, New York: National Bureau of Economic Research.

Beine, M., Docquier, F., and Rapoport, H. (2001): “Brain drain and economic growth: theory and evidence”, Journal of Development Economics, (64): 275-289.

Bhagwati, J.N., Hamada, K. (1974): “The brain drain, international integration of markets for professionals and unemployment”, Journal of Development Economics, 1(1): 19-42.

Blöndal, S., S. Field and N. Girouard (2002), “Investment in human capital through post-compulsory Education and training: Selected Efficiency and Equity Aspects”, OECD Economics Department, Working Papers, No. 333, OECD Publishing.
doi: 10.1787/778845424272

Chen, Zhao, Lu, Ming and Jin, Yu (2004): “Regional differences of human capitals an education development in China: an estimation of the panel data”, World Economics, (12): 25-31. (In Chinese)

Chen, Hung-Ju (2008): “The endogenous probability of migration and economic growth”, Economic Modelling, 25: 1111-1115.

Fan, C.Cindy (2002): “The elite, the natives, and the outsiders: migration and labor market segmentation in urban China”, Annals of the Association of American Geographers, 92(1):103-204.

Fan, Lida (2009): “Measuring Interprovincial Flows of Human Capital in China: 1995–2000”, Popul Res Policy Rev, 28: 367–387.

Fleisher, Belton, Li, Hai-zheng and Zhao, Min-Qiang (2010): “Human capital, economic growth, and regional inequality in China”, Journal of Development Economics, (92), 215-232.

Greene, W.H. (2008): “Econometric Analysis”, 6th edition, Prentice Hall.

Gu, Jia-ning (2006): “the agglomeration effect and status quo of human capital in China”, Market Modernization, (7): 251-254. (In Chinese)

Heckman, James J. (2005): “China’s human capital investment”, China Economic Review, (16): 50–70.

Katz E and Stark O (1987): “International Migration under Asymmetric Information”, Economic Journal,97:718－726.

Li, De-mian (2010): “The drop-out problems in compulsory education in rural areas: from a perspective of sociology”, Jilin Education, (12): 7-8. (In Chinese)

Liu, Xin-jian, Fan, Jun-feng and Xie, Shun-lin(2008): “Regional difference in human capital in four parts of China: based on data from 26 regions”, Statistics and Decision, (20): 87-89.(In Chinese)

Lu, Yuan-quan, Ma, Lei-xin and He, Qian-qian: “Comparison of human capital among 31 regions in China”，Technology and Management,25(5):117—121. (In Chinese)

Miyagiwa, K. (1991): “Scale economies in education and the brain drain problem”, International Economic Review, 32 (3): 743–759.

Mountford, Andrew (1997): “Can a brain drain be good for the growth in the source economy?”, Journal of Development Economics, (53): 287-303

Ray, Debraj (1998): “Development Economics”, 1st ed. Princeton University Press.

Smith, A. (1952) ：“An inquiry into the nature and causes of the wealth of nations”， In R.M. Hutchins & M. J. Adler (Eds.), Great books of the western world: Vol. 39. Adam Smith. Chicago: Encyclopedia Britannica. (Original work published 1776)

Stark, O., Helmenstein, C. and Prskawetz, A. (1998): “Human capital formation, human capital depletion, and migration: a blessing or a curse?”, Economics Letters, 60(3): 363–367.

Stock, J.H. and Watson, M.W. (2007): “Introduction to Econometrics”, 2nd ed. Pearson Education, Inc.

Vidal, J-P., 1998: “The effect of emigration on human capital formation”, Journal of Population Economics, 11(4): 589–600.

Wang, Xue-min (2009): “Applied Multivariate Analysis”, Shanghai: Shanghai University of Finance & Economics Press. (In Chinese)

Wilson, Kathryn (2001): “The determinants of educational attainment: modeling and estimating the human capital model and education production functions”, Southern Economic Journal, 67(3): 518-551.

Xu, Li and Wang, Fen (2006): “Cluster analysis of human capital based on data from 31 regions”, Technology progress and solution, (1): 69-72. (In Chinese)

Yuan, Rong-hua (2005): “Understand the Zhejiang phenomena”, Journal of Zhejiang University, 35(6):52-61. (In Chinese)

Yang, Xue-yi (2010):“The Current Situation, Reason and Countermeasure of Brain Drain of Gansu University Graduates”, Northwest population, 31(1): 113-119. (In Chinese)

Yang, Yang (2009): “Differences in human capital of China”, Master thesis of Northeast Normal University. (In Chinese)

Yang, Yi-min (2009): “The demand and supply of talents in regions of China”, Statistics and information, 24(7):18-22. (In Chinese)

Ying, Long-gen (1999): “China's Changing Regional Disparities during the Reform Period”, Economic Geography, 75(1): 59-70.

Zhang, Dan-dan, Meng,xin (2010): “Labor market impact of large scale internal migration on Chinese urban ‘native’ workers”, discussion paper in IZA DP No.5288.

Zhang, Dan-dan, Meng, xin, and Wang, De-wen (2009): “The dynamic change in wage gap between urban residents and rural migrants in Chinese cities”, Poverty and Economic Research, working paper.

.

Appendix

Appendix 1 Derive the indirect utility function

$$\underset{x,l}{Max} \quad a\ln l + b\ln x + c\ln y$$

$$s.t \quad w(\bar{L} - l) = px$$

The Lagrange Function is: $L = a\ln l + b\ln x + c\ln y + \lambda(w\bar{L} - wl - px)$

The fisrt order conditions are: $\frac{\partial L}{\partial l} = \frac{a}{l} - \lambda w = 0 \qquad (1)$

$$\frac{\partial L}{\partial x} = \frac{b}{x} - \lambda p = 0 \qquad (2)$$

$$\frac{(1)}{(2)} \quad \Rightarrow \frac{x^*}{l^*} = \frac{bw}{ap} \qquad (3)$$

Insert (3) into the budget constraint, we have:

$$w\bar{L} - wl^* - p \cdot \frac{bwl^*}{ap} = 0$$

thus: $l^* = \frac{a\bar{L}}{a+b} \quad ; \quad x^* = \frac{bw\bar{L}}{(a+b)p}$

Insert l^* and x^* into the utility function, we get:

$$v(w, p, \bar{L}, y) = a\ln(\frac{a\bar{L}}{a+b}) + b\ln(\frac{bw\bar{L}}{a+b}) + c\ln y$$

$$\Leftrightarrow \quad v(w, p, \bar{L}, y) = a\ln(a\bar{L}) - a\ln(a+b) + b\ln(\frac{bw\bar{L}}{p}) - b\ln(a+b) + c\ln y$$

$$\Leftrightarrow \quad v(w, p, \bar{L}, y) = a\ln(a\bar{L}) + b\ln(\frac{bw\bar{L}}{p}) + c\ln y - (a+b)\ln(a+b)$$

Appendix 2 Proof of the existence of e*

Since $c(e_i)$ is a monotonic continuous function, and $e_i \in (0, E]$, thus the maximum value of is $c(E)$. If $c(E) \geq \frac{b \ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{(\ln y_A - \ln y_B)}$, then and the right hand side of 5.2.6 must fall into the range of $c(e_i)$.Therefore, there must be an e* which makes 5.2.6 holds when $e_i \geq e^*$. On the contrary, if $c(E) < \frac{b \ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{(\ln y_A - \ln y_B)}$, $c(e_i) \leq c(E) < \frac{b \ln(\frac{W_B \bullet P_A}{W_A \bullet P_B})}{(\ln y_A - \ln y_B)}$, e* will not exist.

Appendix 3 Partial derivatives

$$N(w_A, w_B, p_A, p_B, y_A, y_B, C_A, C_B) = \int_0^{e^*} g(e_i)de_i + (1-\pi_A)\bullet\int_{e^*}^{e^{**}} g(e_i)de_i + (1-\pi_B)\bullet\int_{e^{**}}^{E} g(e_i)de_i$$

$$e^* = C^{-1}(\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)});\quad e^{**} = \varphi^{-1}(\frac{(C_A - C_B)(1+r)}{W_A - W_B})$$

According to Leibniz's Formula, we can compute the partial derivatives as following:

(1) $\frac{\partial N}{\partial y_A}$ and $\frac{\partial N}{\partial y_B}$;

$$\begin{aligned}
\frac{\partial N}{\partial y_A} &= \frac{\partial e^*}{\partial y_A}\cdot g(e^*) + (1-\pi_A)\cdot(\frac{\partial e^{**}}{y_A}\cdot g(e^{**}) - \frac{\partial e^*}{y_A}\cdot g(e^*)) + (1-\pi_B)\cdot(-\frac{\partial e^{**}}{\partial y_A}\cdot g(e^{**})) \\
&= \frac{\partial e^*}{\partial y_A}\cdot g(e^*) - (1-\pi_A)\cdot\frac{\partial e^*}{y_A}\cdot g(e^*) \\
&= \pi_A\cdot g(e^*)\cdot C^{-1\prime}(\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)})\cdot(-\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)^2})\cdot\frac{1}{y_A} \\
&= \underbrace{-\pi_A\cdot g(e^*)}_{<0}\cdot\underbrace{C^{-1\prime}(\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)})}_{>0}\cdot\underbrace{\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)^2\cdot y_A}}_{>0} \\
&< 0
\end{aligned}$$

The second term is positive, because $c(\cdot)$ is an increasing and continuous function, $c^{-1}(\cdot)$ which is the inverse function of $c(\cdot)$, is also an increasing and continuous function.

$\frac{\partial N}{\partial y_B}$ is very similar to $\frac{\partial N}{\partial y_A}$, which is:

$$\frac{\partial N}{\partial y_B} = \underbrace{\pi_A\cdot g(e^*)}_{>0}\cdot\underbrace{C^{-1\prime}(\frac{b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)})}_{>0}\cdot\underbrace{\frac{-b\ln(\frac{W_B\bullet P_A}{W_A\bullet P_B})}{(\ln y_A - \ln y_B)^2\cdot y_B}}_{<0} > 0$$

(2) $\frac{\partial N}{\partial C_A}$ and $\frac{\partial N}{\partial C_B}$;

$$\frac{\partial N}{\partial C_A} = \frac{\partial e^*}{\partial C_A} \cdot g(e^*) + (1-\pi_A) \cdot (\frac{\partial e^{**}}{C_A} \cdot g(e^{**}) - \frac{\partial e^*}{C_A} \cdot g(e^*)) + (1-\pi_B) \cdot (-\frac{\partial e^{**}}{\partial C_A} \cdot g(e^{**}))$$

$$= (1-\pi_A) \cdot \frac{\partial e^{**}}{C_A} \cdot g(e^{**}) - (1-\pi_B) \cdot \frac{\partial e^{**}}{\partial C_A} \cdot g(e^{**})$$

$$= \underbrace{(\pi_B - \pi_A) \cdot g(e^{**})}_{<0} \cdot \underbrace{\varphi^{-1\prime}(\frac{(C_A - C_B)(1+r)}{W_A - W_B})}_{<0} \cdot \underbrace{\frac{(1+r)}{(W_A - W_B)}}_{>0}$$

$$> 0$$

The second term is positive, because $\varphi(\cdot)$ is an decreasing and continuous function, $\varphi^{-1}(\cdot)$ is also an decreasing function, which has a positive first order derivative.

Similarly:

$$\frac{\partial N}{\partial C_B} = \underbrace{(\pi_B - \pi_A) \cdot g(e^{**})}_{<0} \cdot \underbrace{\varphi^{-1\prime}(\frac{(C_A - C_B)(1+r)}{(W_A - W_B)})}_{<0} \cdot \underbrace{\frac{-(1+r)}{(W_A - W_B)}}_{<0} < 0$$

(3) $\frac{\partial N}{\partial P_A}$ and $\frac{\partial N}{\partial P_B}$;

$$\frac{\partial N}{\partial P_A} = \frac{\partial e^*}{\partial P_A} \cdot g(e^*) + (1-\pi_A) \cdot (\frac{\partial e^{**}}{P_A} \cdot g(e^{**}) - \frac{\partial e^*}{P_A} \cdot g(e^*)) + (1-\pi_B) \cdot (-\frac{\partial e^{**}}{\partial P_A} \cdot g(e^{**}))$$

$$= \frac{\partial e^*}{\partial P_A} \cdot g(e^*) - (1-\pi_A) \cdot \frac{\partial e^*}{P_A} \cdot g(e^*)$$

$$= \pi_A \cdot g(e^*) \cdot C^{-1\prime}(\frac{b \ln(\frac{W_B \cdot P_A}{W_A \cdot P_B})}{(\ln y_A - \ln y_B)}) \cdot \frac{b}{(\ln y_A - \ln y_B)} \cdot \frac{W_A \cdot P_B}{W_B \cdot P_A} \cdot \frac{W_B}{W_A \cdot P_B}$$

$$= \underbrace{\pi_A \cdot g(e^*)}_{>0} \cdot \underbrace{C^{-1\prime}(\frac{b \ln(\frac{W_B \cdot P_A}{W_A \cdot P_B})}{(\ln y_A - \ln y_B)})}_{>0} \cdot \underbrace{\frac{b}{(\ln y_A - \ln y_B) \cdot P_A}}_{>0}$$

$$> 0$$

Similarly,

$$\frac{\partial N}{\partial P_B} = \pi_A \cdot g(e^*) \cdot C^{-1\prime}\left(\frac{b\ln\left(\frac{W_B \cdot P_A}{W_A \cdot P_B}\right)}{(\ln y_A - \ln y_B)}\right) \cdot \frac{b}{(\ln y_A - \ln y_B)} \cdot \frac{W_A \cdot P_B}{W_B \cdot P_A} \cdot \frac{W_B \cdot P_A}{W_A} \cdot \frac{-1}{P^2{}_B}$$

$$= \underbrace{\pi_A \cdot g(e^*)}_{>0} \cdot \underbrace{C^{-1\prime}\left(\frac{b\ln\left(\frac{W_B \cdot P_A}{W_A \cdot P_B}\right)}{(\ln y_A - \ln y_B)}\right)}_{>0} \cdot \underbrace{\frac{-b}{(\ln y_A - \ln y_B) \cdot P_B}}_{<0}$$

$$< 0$$

(4) $\frac{\partial N}{\partial W_A}$ and $\frac{\partial N}{\partial W_B}$;

$$\frac{\partial N}{\partial W_A} = \frac{\partial e^*}{\partial W_A} \cdot g(e^*) + (1-\pi_A) \cdot \left(\frac{\partial e^{**}}{W_A} \cdot g(e^{**}) - \frac{\partial e^*}{W_A} \cdot g(e^*)\right) + (1-\pi_B) \cdot \left(-\frac{\partial e^{**}}{\partial W_A} \cdot g(e^{**})\right)$$

$$= \pi_A \cdot \frac{\partial e^*}{\partial W_A} \cdot g(e^*) + (\pi_B - \pi_A) \cdot \frac{\partial e^{**}}{\partial W_A} \cdot g(e^{**})$$

$$= \pi_A \cdot g(e^*) \cdot C^{-1\prime}\left(\frac{b\ln\left(\frac{W_B \cdot P_A}{W_A \cdot P_B}\right)}{(\ln y_A - \ln y_B)}\right) \cdot \frac{b}{(\ln y_A - \ln y_B)} \cdot \frac{W_A \cdot P_B}{W_B \cdot P_A} \cdot \frac{W_B \cdot P_A}{P_B} \cdot \frac{-1}{W^2{}_A}$$

$$+(\pi_B - \pi_A) \cdot \varphi^{-1\prime}\left(\frac{(C_A - C_B)(1+r)}{(W_A - W_B)}\right) \cdot (C_A - C_B)(1+r) \cdot \frac{-1}{(W_A - W_B)^2}$$

$$= \underbrace{\pi_A \cdot g(e^*)}_{>0} \cdot \underbrace{C^{-1\prime}\left(\frac{b\ln\left(\frac{W_B \cdot P_A}{W_A \cdot P_B}\right)}{(\ln y_A - \ln y_B)}\right)}_{>0} \cdot \underbrace{\frac{-b}{(\ln y_A - \ln y_B) \cdot W_A}}_{<0} + \underbrace{(\pi_B - \pi_A)}_{<0} \cdot \underbrace{\varphi^{-1\prime}\left(\frac{(C_A - C_B)(1+r)}{(W_A - W_B)}\right)}_{<0} \cdot \underbrace{\frac{-(C_A - C_B)(1+r)}{(W_A - W_B)^2}}_{<0}$$

$$< 0$$

Similarly,

$$\frac{\partial N}{\partial W_B} = \underbrace{\pi_A \cdot g(e^*)}_{>0} \cdot \underbrace{C^{-1\prime}\left(\frac{b\ln\left(\frac{W_B \cdot P_A}{W_A \cdot P_B}\right)}{(\ln y_A - \ln y_B)}\right)}_{>0} \cdot \underbrace{\frac{b}{(\ln y_A - \ln y_B) \cdot W_B}}_{>0} + \underbrace{(\pi_B - \pi_A)}_{<0} \cdot \underbrace{\varphi^{-1\prime}\left(\frac{(C_A - C_B)(1+r)}{(W_A - W_B)}\right)}_{<0} \cdot \underbrace{\frac{(C_A - C_B)(1+r)}{(W_A - W_B)^2}}_{>0}$$

$$> 0$$